2 Public-Health-Methoden

Als Forschungsmethoden bezeichnet man Verfahren und Techniken, die zur Klärung von wissenschaftlichen Fragestellungen dienen. Da es sich bei *Public Health* um ein interdisziplinäres Fach handelt, kann die Herangehensweise an bestimmte Fragestellungen sehr unterschiedlich. Die einzelnen Disziplinen, die sich in Public Health zusammenfinden, bringen jeweils ihre eigenen Methoden mit. Diese können daher sowohl aus dem Bereich der Naturwissenschaften als auch aus dem Bereich der Sozial- und Geisteswissenschaften stammen.
In diesem Kapitel gehen wir zuerst auf eine der Kernwissenschaften von Public Health ein, die *Epidemiologie*. Wir betrachten ihre Rolle in Public Health, schauen uns verschiedene epidemiologische Verfahren zum Messen und Vergleichen an und erläutern spezifische epidemiologische Grundbegriffe wie Expositionen, Outcomes, Validität und Reliabilität. Anschließend werden epidemiologische und klinische Studientypen sowie systematische Übersichtsarbeiten und Metaanalysen vorgestellt. Um die Qualität epidemiologischer Studien beurteilen zu können, ist es von großer Bedeutung, die wichtigsten Fehlerquellen zu kennen.
Der Abschnitt *Demografie* beschäftigt sich mit den Kennziffern zur Beschreibung einer Bevölkerung, z. B. dem Geburtenüberschuss, dem Wanderungssaldo, verschiedenen Sterberaten, der Lebenserwartung und potentiell verlorenen Lebensjahren. Außerdem zeigt er häufig verwendete grafische Darstellungen, z. B. zur Altersstruktur einer Bevölkerung.
Die *Biostatistik* erläutert, wie man trotz vorhandener statistischer Unsicherheit möglichst wahrheitsgemäße Schlussfolgerungen über Populationen und Patientengruppen ziehen kann. Hierzu werden Daten zuerst klassifiziert und transparent zusammengefasst. Begriffe wie Stichprobenvariabilität, Normalverteilung und 95 %-Vertrauensintervall, p-Wert und statistische Signifikanz werden erklärt.
Der Abschnitt *Sozialwissenschaftliche Methoden der Datenerhebung* beschäftigt sich mit der Fragebogenerstellung, insbesondere mit der Formulierung von guten Fragen und möglichen Antworten. Anschließend werden quantitative und qualitative Methoden der Datenerhebung betrachtet und ihre Vor- und Nachteile diskutiert.
Der Abschnitt *Gesundheitsökonomie* stellt zuerst die zentralen gesundheitsökonomischen Studientypen vor. Dabei wird auch die Frage diskutiert, wie man den Nutzen medizinischer Maßnahmen quantifizieren kann. Anschließend wird erläutert, wie man Ergebnisse gesundheitsökonomischer Studien ausdrücken und interpretieren kann. Der Abschnitt schließt mit der für die Gesundheitsökonomie zentralen Frage: Wie viele Ressourcen wollen Gesellschaften für zusätzliche Gesundheit aufwenden?

Schweizerische Lernziele: CPH 1, CPH 5–20, CPH 27

2.1 Epidemiologie

Oliver Razum, Patrick Brzoska, Matthias Egger

Die Epidemiologie ist eine Kernwissenschaft für Public Health: Sie ist unentbehrlich, um den Gesundheitszustand auf der Bevölkerungsebene zu beschreiben, Krankheitsursachen und damit Interventionsmöglichkeiten zu identifizieren und deren Wirksamkeit zu messen. Wörtlich übersetzt ist Epidemiologie die Lehre von dem, was „über das Volk kommt“ [von *epi* (gr.): über und *démos* (gr.): Volk]. Sie untersucht die Verteilung von Krankheiten, Todesfällen und anderen gesundheitlichen Ereignissen

https://doi.org/10.1515/9783110466867-006

(„Outcomes“) in Bevölkerungen oder Bevölkerungsgruppen, aber auch von Risikofaktoren und schützenden Faktoren (beide werden unter dem Begriff „Expositionen“ zusammengefasst). Die deskriptive Epidemiologie beschreibt dabei die Verteilung von Outcomes und Expositionen, die analytische Epidemiologie schließt aus den Verteilungsmustern auf mögliche Krankheitsursachen und setzt dazu epidemiologische Studiendesigns wie Kohortenstudien und Fall-Kontroll-Studien ein. Bei der Betrachtung der Studienergebnisse stellen EpidemiologInnen systematische Überlegungen zu möglichen Verzerrungen und ihren Folgen sowie zur Ursächlichkeit (Kausalität) der beobachteten Zusammenhänge an. Die Ergebnisse solcher epidemiologischer Studien helfen, präventive Interventionsmaßnahmen zu erarbeiten und diese zu evaluieren.

2.1.1 Die Rolle der Epidemiologie in Public Health

Epidemiologie – Definition und Überblick

Die Epidemiologie untersucht die Verteilung von gesundheitsrelevanten Ereignissen und Determinanten in Bevölkerungen oder Bevölkerungsgruppen. Solche *gesundheitsrelevanten Ereignisse* – „Outcomes“ – sind v. a. Krankheiten und Todesfälle. Zu den *Determinanten* gehören Risikofaktoren und schützende (*protektive*) Faktoren – EpidemiologInnen sprechen hier allgemein von „Expositionen“. Expositionen können sich aus dem individuellen Verhalten von Menschen ergeben (z. B. Rauchen oder regelmäßiger körperlicher Aktivität), aber auch aus der physikalischen Umwelt (z. B. Zugang zu Gesundheitsdiensten). Mit einer „Bevölkerung“ oder Population im epidemiologischen Sinne können – je nach Situation – alle Menschen eines Landes gemeint sein, aber auch Untergruppen wie z. B. alle Menschen über 65 Jahre oder alle TeilnehmerInnen einer Studie.

Untersuchungsgegenstand der Epidemiologie sind heute nicht nur Infektionskrankheiten, sondern auch nichtübertragbare, chronische Erkrankungen wie etwa der Diabetes mellitus und seine Risikofaktoren. Epidemiologie beschäftigt sich darüber hinaus z. B. auch mit berufsbedingten Erkrankungen und Unfällen. So konnten EpidemiologInnen das gehäufte Auftreten von Blasenkrebs nach einer beruflichen Exposition gegenüber aromatischen Aminen nachweisen. Ein weiteres Beispiel ist der Nachweis der Häufung von Verkehrsunfällen unter jungen männlichen Autofahrern jeweils in der Nacht von Freitag und Samstag.

Solche Kenntnisse über die Verteilung von Gesundheitsproblemen und ihren Risikofaktoren in der Bevölkerung ermöglichen es, die jeweilige Größe der Probleme quantitativ zu beschreiben. Hieraus sind dann Rückschlüsse auf die Krankheitsursachen möglich, sodass geeignete Maßnahmen zur Prävention definiert werden können. Schließlich kann auch die Wirksamkeit solcher Maßnahmen evaluiert werden. EpidemiologInnen wenden hierbei deskriptive und analytische Verfahren an.

Die **deskriptive Epidemiologie** beschreibt ein Gesundheitsproblem, indem sie die folgenden Fragen beantwortet:

- Wann treten die Krankheits-/Todesfälle auf? (Verteilung über die Zeit)
- Wo treten die Krankheits-/Todesfälle auf? (geografische Verteilung)
- Wer ist erkrankt? Wie viele Menschen erkranken/versterben? Wer ist exponiert? Wie viele Menschen sind exponiert?

Die Fragen *Wann?*, *Wo?* und *Wer?* (*Time, Place, Person*) werden als die drei epidemiologischen Fragen bezeichnet. Sie sind die Grundlage des epidemiologischen Arbeitens. Eine Form der deskriptiven Epidemiologie und gleichzeitig Datenquelle für weitere epidemiologische Auswertungen ist die *Gesundheitsberichterstattung* (GBE). Sie umfasst z. B. die Datensätze der Todesursachenstatistik. Diese enthält u. a. Angaben zur Anzahl der Todesfälle, unterschieden (*stratifiziert*) nach Todesursachen, Alter, Geschlecht und Sterbejahr. Details zur GBE finden sich jeweils auf den Websites von *Statistik Schweiz, Statistik Austria* und der *Gesundheitsberichterstattung des Bundes* in Deutschland (s. Internet-Ressourcen).

Die **analytische Epidemiologie** befasst sich mit der Ermittlung von Risikofaktoren und von Krankheitsursachen. Dazu werden epidemiologische Studiendesigns eingesetzt, bei denen Vergleiche zwischen Populationen angestellt werden. Auch in der analytischen Epidemiologie werden drei Fragen beantwortet (s. a. Kap. 2.1.3 und 2.1.8):

- Besteht eine Assoziation zwischen einem vermuteten Risikofaktor und dem untersuchten Outcome?
- Wie stark ist die Assoziation?
- Ist die beobachtete Assoziation ursächlich (kausal)?

Sind auf diese Weise Risikofaktoren identifiziert, die zum Auftreten des untersuchten Outcomes beitragen (*Attributables Risiko*, s. Kap. 2.1.3), so können geeignete präventive Interventionen entwickelt werden. Deren Wirksamkeit müsste sich durch eine verringerte Häufigkeit des Outcomes zeigen lassen, etwa mit Hilfe von Daten der GBE. In der Realität treten dabei jedoch häufig *Störfaktoren* auf (s. Kap. 2.1.8). Ein wissenschaftlich solider Wirksamkeitsnachweis auf höchstem Evidenzniveau kann nur experimentell durch eine *randomisierte kontrollierte Studie* erbracht werden (s. Kap. 2.1.6).

Zusätzlich zur Unterteilung in deskriptive und analytische Epidemiologie werden innerhalb der Epidemiologie noch verschiedene Themen- und Forschungsbereiche unterschieden. Beispiele hierfür sind die Umweltepidemiologie, die Ernährungsepidemiologie, die Sozialepidemiologie, die klinische Epidemiologie und die molekulare Epidemiologie.

Einige Meilensteine der Epidemiologie

Das „Denken in Bevölkerungen" ist keine neue Erfindung. Wichtige Grundprinzipien der Epidemiologie sind schon seit Jahrzehnten oder Jahrhunderten bekannt. Die folgenden Pioniere der Epidemiologie lassen wichtige Ideen und Herangehensweisen erkennen, die in der Epidemiologie eine große Rolle spielen:

John Graunt (1620–1674), ein englischer Kaufmann, analysierte die Listen aller Todesfälle, die schon damals in London geführt wurden – ähnlich, wie dies in der *Todesursachenstatistik* heute noch geschieht. Graunt stellte fest, dass Kinder ein höheres Sterberisiko als Erwachsene hatten, dass das Sterberisiko bei Männern höher war als bei Frauen, und dass die Sterblichkeit in London höher lag als auf dem Lande. Hieraus schlussfolgerte er, dass die Risiken für Krankheit und Tod nicht zufällig und nicht gleichmäßig in der Bevölkerung verteilt sind. Diese Erkenntnis mag banal erscheinen, sie ist aber Grundlage jeglicher Epidemiologie. Würden Krankheiten und Todesfälle rein zufällig auftreten, so könnte man keine Risikofaktoren identifizieren (wie z. B. das Rauchen als Risikofaktor für Lungenkrebs) oder Bevölkerungsgruppen benennen, die ein erhöhtes Risiko aufweisen (wie etwa allein stehende, ältere Männer für Suizid).

Der englische Arzt **John Snow** (1813–1858) untersuchte die großen *Cholera-Ausbrüche*, die im 19. Jahrhundert in den Städten zu Tausenden von Toten führten. Damals waren Erreger und Übertragungsweg der Seuche noch unbekannt. Snow zeigte mit deskriptiven und analytischen epidemiologischen Methoden, dass kontaminiertes Trinkwasser eine wesentliche Rolle bei der Übertragung der Cholera in London spielte. Viele Jahre vor der Kultivierung des Erregers *Vibrio cholerae* durch *Robert Koch* konnte er aus seinen Studienergebnissen wirksame Präventionsmaßnahmen ableiten.

Der deutsche Mediziner und Begründer der Zellpathologie **Rudolf Virchow** (1821–1902) leistete Pionierarbeit auf dem Gebiet der Sozialepidemiologie. Virchow beobachtete während einer *Hungertyphus-Epidemie* (Typhus exanthematicus; Läusefleckfieber) in Oberschlesien, dass Armut krank macht und Krankheit somit auch gesellschaftliche Ursachen hat. Medikamente allein reichen nicht, um den Gesundheitszustand der Bevölkerung zu verbessern, solange es an bezahlter Arbeit, Bildung und sozialer Absicherung mangelt. Virchow prägte 1848 den Satz „Die Medizin ist eine soziale Wissenschaft, und die Politik ist weiter nichts als Medizin im Großen".

Der Epidemiologe **Richard Doll** (1912–2005) und der Statistiker **Austin Bradford Hill** (1897–1991) führten die *British Doctors Study* durch, eine modellhafte Kohortenstudie (s. Kap. 2.1.5) zum Einfluss des Rauchens auf die Sterblichkeit an Lungenkrebs und anderen Erkrankungen. Im Jahr 1951 rekrutierten Doll und Hill mehr als 34.000 Ärzte aus dem britischen Ärzteregister und fragten nach deren Rauchgewohnheiten. Sie beobachteten die Ärzte über viele Jahre und verglichen unter anderem die Häufigkeit von Todesfällen an Lungenkrebs und Herz-Kreislauf-Krankheiten unter exponierten und nicht exponierten Ärzten. Beide trugen dazu bei, dass *Rauchen als Risikofaktor für Lungenkrebs* erkannt wurde, lange bevor die Mechanismen der Krebsentstehung

auf zellulärer Ebene verstanden wurden. Hill war darüber hinaus einer der Pioniere auf dem Gebiet der *randomisierten Studien* (s. Kap. 2.1.6) und entwickelte die nach ihm benannten *Bradford-Hill-Kriterien für Kausalität* (s. Kap. 2.1.7).

2.1.2 Epidemiologische Verfahren zum Messen und Vergleichen

Häufigkeitsmaße für Expositionen und Outcomes

Zur Untersuchung der Häufigkeit von *Expositionen* und *Outcomes* nutzt die Epidemiologie verschiedene deskriptive Maßzahlen.

Absolute Zahl: Die grundlegendste deskriptive Maßzahl ist die absolute Zahl. Sie gibt die Anzahl von Personen an, die einer bestimmten Exposition ausgesetzt sind oder einen bestimmten Outcome aufweisen. Die absolute Zahl ist eine wichtige Grundlage der Gesundheitsberichterstattung und wird aus unterschiedlichen routinemäßig erhobenen Daten, amtlichen Statistiken oder epidemiologischen *Surveys* gewonnen. So zeigt z. B. die amtliche Pflegestatistik, dass am 15. Dezember 2013 die absolute Zahl an Menschen, die nach der Definition des *Sozialgesetzbuchs XI* in Nordrhein-Westfalen pflegebedürftig waren, 581.492 betrug. In Hessen waren zum gleichen Zeitpunkt 205.126 Menschen pflegebedürftig.

Prävalenz: Ein Nachteil absoluter Zahlen ist, dass sie keinen Vergleich zwischen einzelnen Regionen oder verschiedenen Zeitpunkten erlauben. So kommt die höhere Zahl der Pflegebedürftigen im deutschen Bundesland Nordrhein-Westfalen vermutlich dadurch zustande, dass die Bevölkerung dort größer ist als im Bundesland Hessen. Ein Vergleich der absoluten Zahlen (*Zähler*) wird daher erst dann möglich, wenn sie in Bezug zur Bevölkerungsgröße (*Nenner*) der jeweiligen Regionen gesetzt werden. (Auf mögliche Altersstruktureffekte wollen wir an dieser Stelle nicht eingehen [s. hierzu Kap. 2.2.2].)

Die daraus resultierende Maßzahl heißt **Punktprävalenz**. Sie beschreibt den Anteil der Exponierten bzw. den Anteil derjenigen, die einen bestimmten Outcome aufweisen (zum Beispiel pflegebedürftig sind), jeweils zu einem definierten Zeitpunkt. Dieser Anteil wird oft pro 100 Personen, manchmal auch pro 1.000, 10.000 oder 100.000 Personen der Gesamtbevölkerung angegeben:

$$\text{Punktprävalenz} = \frac{\text{Personen m. Exposition bzw. Outcome zu einem definierten Zeitpunkt}}{\text{Gesamtbevölkerung zum gleichen Zeitpunkt}} \; (\cdot 100)$$

Die Punktprävalenz von Pflegebedürftigkeit am 15. Dezember 2013 betrug damit in Nordrhein-Westfalen, wo zu diesem Zeitpunkt 17.571.856 Menschen lebten,

$$581.492 \,/\, 17.571.856 \cdot 100 = 3{,}3 \text{ pro } 100 \text{ Personen.}$$

Insgesamt waren also 3,3 % der Menschen in Nordrhein-Westfalen pflegebedürftig. In Hessen lag die Punktprävalenz zum gleichen Zeitpunkt bei 3,4 %:

$$205.126 / 6.045.425 \cdot 100 = 3{,}4 \text{ pro } 100 \text{ Personen}$$

Da sich die Punktprävalenz lediglich auf einen einzigen Zeitpunkt bezieht, stellt sie eine Momentaufnahme dar. Sie ist daher nicht als Maßzahl für Erkrankungen mit einer kurzen Dauer (z. B. Durchfallerkrankungen oder Erkältungen) geeignet. In solchen Fällen wird als alternatives Prävalenzmaß die **Periodenprävalenz** berechnet, die sich auf einen Zeitraum (etwa einen Monat oder ein Jahr) bezieht. Im Zähler der Periodenprävalenz befinden sich alle Fälle zu Beginn des betrachteten Zeitraums sowie alle in diesem Zeitraum neu aufgetretenen Fälle:

$$\text{Periodenprävalenz} = \frac{\text{Erkrankte zu Beginn eines Zeitraums + Neuerkrankte im Zeitraum}}{\text{Mittlere Bevölkerung im Zeitraum}} \; (\cdot 100)$$

Als mittlere Bevölkerung im Zeitraum wird in der Regel der Durchschnitt aus der Bevölkerungszahl zu Beginn und zum Ende des Betrachtungszeitraums angegeben.

Einige epidemiologische Lehrbücher verwenden bei der Formel der Punkt- und Periodenprävalenz im Nenner nur die Bevölkerung „unter Risiko“ (*at risk*). Die Bezugsbevölkerung besteht in diesem Fall nur aus denjenigen, die den Outcome ausbilden können. Eine solche Definition ist z. B. bei Fragestellungen sinnvoll, die sich mit geschlechtsspezifischen Krankheiten wie Prostatakrebs oder Gebärmutterhalskrebs beschäftigen.

Inzidenz: Die Periodenprävalenz ist eine statische Maßzahl. Die Inzidenz erlaubt es hingegen, Veränderungen innerhalb eines Zeitraums abzubilden. Das geschieht, indem der Zähler nur die neu aufgetretenen (*inzidenten*) Fälle eines bestimmten Zeitraums berücksichtigt. In der Regel wird die Inzidenz nur für Outcomes (z. B. Erkrankungen), seltener für Expositionen verwendet. In der Epidemiologie lassen sich verschiedene Inzidenzmaße unterscheiden.

Die **kumulative Inzidenz** (auch als *Inzidenzrisiko* oder nur kurz als *Risiko* bezeichnet) ist die Wahrscheinlichkeit, mit der eine Person in einem bestimmten Zeitraum erkrankt. Sie ist das Verhältnis der Zahl an Neuerkrankungen in einem definierten Zeitraum zur Zahl der Bevölkerung unter Risiko zu Beginn des Zeitraums und wird meist pro 1.000 oder 100.000 Personen angegeben.

$$\text{kumulative Inzidenz} = \frac{\text{Neuerkrankte in einem definierten Zeitraum}}{\text{Bevölkerung unter Risiko zu Beginn des Zeitraums}}$$

Die kumulative Inzidenz ist eine geeignete Maßzahl, wenn Veränderungen innerhalb der Bevölkerung unter Risiko (durch Zu- und Abwanderungen sowie durch Gebur-

ten und Sterbefälle) vernachlässigt werden können. Da dies bei vielen epidemiologischen Fragestellungen jedoch nicht der Fall ist und die Bevölkerung unter Risiko im definierten Zeitraum genauer abgebildet werden muss, wird statt der kumulativen Inzidenz die **Inzidenzrate** berechnet. Auch hierbei werden Neuerkrankte in einem definierten Zeitraum betrachtet. Im Nenner befindet sich dann aber die mittlere Bevölkerung unter Risiko:

$$\text{Inzidenzrate} = \frac{\text{Neuerkrankte in einem definierten Zeitraum}}{\text{Mittlere Bevölkerung unter Risiko im gleichen Zeitraum}}$$

Die *mittlere Bevölkerung unter Risiko* ist meist der Durchschnitt aus der Bevölkerung unter Risiko zu Beginn und zum Ende des Betrachtungszeitraums. Wie die kumulative Inzidenz wird die Inzidenzrate oft pro 1.000 oder 100.000 Personen angegeben. Im Jahr 2012 erkrankten z. B. in Deutschland 10.398 Männer an schwarzem Hautkrebs (da es keine flächendeckende Krebsregistrierung gibt, handelt es sich hierbei um einen Schätzwert). Die mittlere männliche Bevölkerung umfasste in diesem Jahr 39.305.462 Personen. Hieraus lässt sich eine Inzidenzrate von (10.398/ 39.305.462) · 100.000 = 26,5 pro 100.000 Männern errechnen. Dies bedeutet, dass im Jahr 2012 in Deutschland pro 100.000 männliche Personen etwa 27 Männer neu an schwarzem Hautkrebs (Melanom) erkrankten.

In epidemiologischen Studien setzt ein solches Vorgehen voraus, dass alle Studienteilnehmer zeitgleich in die Studie aufgenommen werden. In der Praxis ist das oft nicht möglich, da die Rekrutierung meist einen längeren Zeitraum beansprucht. Außerdem nehmen nicht alle Personen, die zu Beginn in eine Studie eingeschlossen werden, auch bis zum Ende daran teil. Manche von ihnen entwickeln den Outcome (d. h. sie erkranken), andere Teilnehmer wollen nicht länger an der Studie teilnehmen, oder man weiß nichts mehr über ihren Verbleib. Sie alle scheiden aus der Studie aus und zählen nicht länger zur Bevölkerung unter Risiko. Um die unterschiedlichen Zeiträume zu berücksichtigen, die die Personen zur Bevölkerung unter Risiko gehören, wird in epidemiologischen Studien bei der Berechnung der Inzidenzrate im Nenner statt der mittleren Bevölkerung häufig die *Personenzeit unter Risiko* verwendet. In diesem Fall heißt die Inzidenzrate auch **Inzidenzdichte**:

$$\text{Inzidenzdichte} = \frac{\text{Neuerkrankte in einem definierten Zeitraum}}{\text{Personenzeit unter Risiko im gleichen Zeitraum}}$$

Sie wird meist pro 1.000 oder 100.000 Personenjahre angegeben.

Spezielle Inzidenzmaße für die Untersuchung der Sterblichkeit sind die Mortalität und die Letalität (s. a. Kap. 2.2.3). Die **Mortalität** (auch Mortalitätsrate oder Sterbeziffer) bezeichnet die Zahl der Gestorbenen in einem bestimmten Zeitraum (*inzidente Sterbefälle*) im Verhältnis zur mittleren Bevölkerung in dieser Zeit. Die Mortalität ist

eine Maßzahl, die häufig in der Gesundheitsberichterstattung verwendet und meistens pro 100.000 Personen der Bevölkerung angegeben wird:

$$\text{Mortalitätsrate} = \frac{\text{Gestorbene innerhalb eines Zeitraums}}{\text{Mittlere Bevölkerung unter Risiko im gleichen Zeitraum}} \ (\cdot 100.000)$$

Im Jahr 2014 sind in Deutschland z. B. 223.758 Menschen an bösartigen Tumoren gestorben. Die mittlere Bevölkerung betrug im gleichen Jahr 80.982.500 Personen. Daraus lässt sich eine Mortalitätsrate von (223.758 / 80.982.500) · 100.000 = 276,3 pro 100.000 Personen errechnen. Im Jahr 2014 starben in Deutschland also etwa 276 von 100.000 Menschen an bösartigen Tumoren.

Die **Letalität** wird meist in Prozent ausgedrückt und bezeichnet die Zahl der in einem definierten Zeitraum an einer Krankheit Gestorbenen im Verhältnis zur Zahl der Erkrankten im gleichen Zeitraum:

$$\text{Letalität} = \frac{\text{An einer Krankheit Gestorbene innerhalb eines Zeitraums}}{\text{Alle von der Krankheit betroffenen Personen im Zeitraum}} \ (\cdot 100)$$

Beispiel: Im Jahr 1976 erkrankten 182 Veteranen der *American Legion*, die sich in Philadelphia zu einem Treffen versammelt hatten, an einer bis dahin unbekannten Form der Lungenentzündung. Neunundzwanzig der 182 erkrankten Personen verstarben (Letalität 16 %). Der Erreger erhielt den Namen *Legionella pneumophila*, und die Krankheit wurde fortan als Legionärskrankheit bekannt.

2.1.3 Assoziationsmaße für Expositionen und Outcomes

Aufgabe der Epidemiologie ist es nicht nur, die Häufigkeit von Expositionen und Outcomes in einer Bevölkerung zu untersuchen, sondern auch die Stärke des Zusammenhangs (*Assoziation*) zwischen ihnen zu bestimmen. Dafür stehen unterschiedliche **Assoziationsmaße** zur Verfügung. Um einen solchen Zusammenhang zu berechnen, nutzt die Epidemiologie häufig die sogenannte *Vier-Felder-Tafel* als Hilfsmittel. Ein hier festgestellter Zusammenhang sagt jedoch noch nichts über eine mögliche Ursache-Wirkungs-Beziehung aus (s. Kap. 2.1.8).

Vier-Felder-Tafel: Eine Vier-Felder-Tafel (*Kontingenztafel*) stellt die Häufigkeit einer Exposition in Abhängigkeit zu einem Outcome dar (s. Tab. 2.1).

Tab. 2.1: Gerüst einer Vier-Felder-Tafel.

		Outcome		
		Ja	*Nein*	*Summe*
Exposition	*Ja*	a	b	a+b
	Nein	c	d	c+d
	Summe	a+c	b+d	a+b+c+d

Die vier Felder bezeichnen hierbei:
- die exponierten Personen, bei denen der Outcome aufgetreten ist (a),
- die exponierten Personen, bei denen der Outcome nicht aufgetreten ist (b),
- die nicht exponierten Personen, bei denen der Outcome aufgetreten ist (c),
- die nicht exponierten Personen, bei denen der Outcome nicht aufgetreten ist (d).

Weitere wichtige Informationen der Vier-Felder-Tafel ergeben sich aus den fünf Randsummen. Sie geben Auskunft über
- alle exponierten Personen (a+b),
- alle nicht exponierten Personen (c+d),
- alle Personen, bei denen der Outcome aufgetreten ist (a+c),
- alle Personen, bei denen der Outcome nicht aufgetreten ist (b+d),
- alle Personen, die untersucht wurden (a+b+c+d).

Relatives Risiko und Relative Rate: Aus der Vier-Felder-Tafel können die *kumulativen Inzidenzen* unter den exponierten Personen (a/[a+b]) und den nicht exponierten Personen (c/[c+d]) abgelesen werden. Das Verhältnis der beiden Inzidenzen ist ein Maß für die Stärke des Zusammenhangs zwischen Exposition und Outcome. Es wird als **Relatives Risiko (RR)** bezeichnet:

$$\text{Relatives Risiko} = \frac{\text{Kumulative Inzidenz unter den Exponierten}}{\text{Kumulative Inzidenz unter den Nichtexponierten}} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Ist das Relative Risiko größer als 1, weist das auf ein höheres Risiko der Exponierten hin, den Outcome auszubilden (z. B. zu erkranken). Bei einem Relativen Risiko kleiner als 1 ist das Risiko für die Exponierten, den Outcome auszubilden, geringer als unter den Nichtexponierten. Je weiter das Relative Risiko von 1 entfernt ist, desto größer ist der Unterschied zwischen Exponierten und Nichtexponierten und damit auch der Zusammenhang zwischen Exposition und Outcome.

Liegen statt kumulativer Inzidenzen *Inzidenzraten* vor, ist es strenggenommen nicht korrekt, von einem Relativen Risiko zu sprechen, da man Zähler und Nenner nicht als Wahrscheinlichkeiten interpretieren kann. Der Quotient aus der Inzidenz-

rate unter den Exponierten und der Inzidenzrate unter den Nichtexponierten wird stattdessen als **Relative Rate** bezeichnet. Er wird wie das Relative Risiko berechnet:

$$\text{Relative Rate} = \frac{\text{Inzidenzrate unter den Exponierten}}{\text{Inzidenzrate unter den Nichtexponierten}}$$

In der *British Doctors Study* betrug z. B. nach 20 Jahren Beobachtungszeit die Inzidenzdichte für Speiseröhrenkrebs 14 pro 100.000 bei Zigarettenrauchern und 3 pro 100.000 bei Nichtrauchern. Die relative Rate berechnet sich folgendermaßen:

$$\text{Relative Rate} = (14 / 100.000) / (3 / 100.000) = 4{,}7$$

Raucher haben demnach eine fast fünfmal so hohe Rate wie Nichtraucher, an Speiseröhrenkrebs zu versterben.

Um Relative Risiken oder Relative Raten berechnen zu können, muss die kumulative Inzidenz bzw. die Inzidenzrate unter den Exponierten und Nichtexponierten bekannt sein. In Kohorten- und randomisierten kontrollierten Studien (s. Kap. 2.1.5 und Kap. 2.1.6) ist dies der Fall. Bei Studientypen, wo dies nicht möglich ist (z. B. Querschnitt- oder Fall-Kontroll-Studien), wird stattdessen als alternatives Assoziationsmaß die Odds Ratio verwendet.

Odds Ratio: Die **Odds Ratio (OR)** basiert – anders als das Relative Risiko – nicht auf der kumulativen Inzidenz, sondern auf der Chance (im Englischen als Odds bezeichnet) von Exponierten und Nichtexponierten, dass ein Outcome auftritt. Die Chance ist hierbei ein Quotient, bestehend aus der Wahrscheinlichkeit für einen Outcome und seiner Gegenwahrscheinlichkeit:

$$\text{Odds} = \frac{\text{Wahrscheinlichkeit für einen Outcome}}{1 - (\text{Wahrscheinlichkeit für einen Outcome})}$$

Die Odds Ratio setzt also die Chancen von Exponierten und Nichtexponierten zueinander ins Verhältnis. Sie ist damit ein Quotient zweier Quotienten und wird deshalb manchmal auch als *Quotenquotient* oder *Chancenverhältnis* bezeichnet. Die beiden Chancen lassen sich leicht aus einer Vier-Felder-Tafel ablesen (a/b und c/d). Durch die Kehrwertregel kann man die Formel für die Odds Ratio schließlich wie hier dargestellt vereinfachen:

$$\text{Odds Ratio} = \frac{\text{Odds unter den Exponierten}}{\text{Odds unter den Nichtexponierten}} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c}$$

Interpretieren lässt sich die Odds Ratio ähnlich wie das Relative Risiko oder die Relative Rate. Werte über 1 weisen auf eine höhere Chance, Werte unter 1 auf eine

geringere Chance der Exponierten hin, den Outcome auszubilden. Je weiter eine Odds Ratio nach unten oder oben von der 1 abweicht, desto stärker ist der Zusammenhang zwischen Exposition und Outcome. Ist der Outcome selten, liegen die Odds Ratio und das Relative Risiko eng beieinander. Wenn der Outcome hingegen häufig ist, können Odds Ratio und Relatives Risiko stark voneinander abweichen.

Die Vier-Felder-Tafel in Tab. 2.2 illustriert die Berechnung und Interpretation der Odds Ratio am Beispiel einer Fall-Kontroll-Studie zum Einfluss von Neuroleptika (*Exposition*) auf die Entstehung einer venösen Thromboembolie (VTE; *Outcome*).

Tab. 2.2: Vier-Felder-Tafel zum Zusammenhang von Neuroleptikaeinnahme und venöser Thromboembolie (VTE).

		Outcome (VTE)					
		Ja		*Nein*		*Summe*	
Exposition (Neuroleptikaeinnahme)	*Ja*	2.126	a	4.752	b	6.878	a+b
	Nein	23.406	c	84.739	d	108.145	c+d
	Summe	25.532	a+c	89.491	b+d	115.023	a+b+c+d

Odds Ratio = (2.126 · 84.739) / (4.752 · 23.406) = 1,62

Quelle der Originaldaten: Parker C, Coupland C, Hippisley-Cox J. Antipsychotic drugs and risk of venous thromboembolism: nested case-control study. British Medical Journal 2010; 341: c4245.

Die Autoren untersuchten insgesamt 115.023 Personen, die in Hausarztpraxen in Großbritannien registriert waren. Hiervon entwickelten 25.532 PatientInnen zwischen 1996 und 2007 eine VTE (*Fälle*), bei den übrigen 89.491 Personen kam es nicht zu einer VTE (*Kontrollen*). Insgesamt nahmen 6.878 Personen Neuroleptika ein (*Exponierte*), 108.145 Personen taten dies nicht (*Nichtexponierte*). Von den Neuroleptika-NutzerInnen litten 2.126 an einer VTE, 4.752 nicht. Die Odds einer VTE betrug daher unter den Exponierten 2.126 / 4.752. Von denjenigen, die keine Neuroleptika einnahmen, litten 23.406 an einer VTE. Ihre Odds lag damit bei 23.406 / 84.739. Um hieraus die Odds Ratio zu errechnen, wird ein Quotient aus beiden Odds gebildet:

$$OR = (2.126 / 4.752) / (23.406 / 84.739) = (2.126 \cdot 84.739) / (4.752 \cdot 23.406) = 1{,}6$$

Demnach ist also die Chance bei Personen, die Neuroleptika einnehmen, eine VTE zu bekommen, 1,6-mal so hoch wie bei Personen, die keine Neuroleptika einnehmen.

Attributables Risiko: Besteht eine Ursache-Wirkungs-Beziehung zwischen Exposition und Outcome (s. Kap. 2.1.8), erlaubt das **attributable Risiko (AR)** den Anteil bei den neuen Erkrankungen zu ermitteln, der auf die Exposition zurückzuführen ist. Das attributable Risiko wird meist in Prozent angeben und wie folgt berechnet:

$$\text{Attributables Risiko} = \frac{\left(\begin{array}{c}\text{Kum. Inzidenz unter}\\ \text{den Exponierten}\end{array}\right) - \left(\begin{array}{c}\text{Kum. Inzidenz unter den}\\ \text{Nichtexponierten}\end{array}\right)}{\text{Kum. Inzidenz unter den Exponierten}} \; (\cdot 100)$$

In der oben erwähnten Analyse der *British Doctors Study* lässt sich das attributable Risiko zum Zusammenhang von Rauchen und Speiseröhrenkrebs-Mortalität folgendermaßen berechnen:

$$(14 / 100.000 - 3 / 100.000) / (14 / 100.000) \cdot 100 = 78{,}6\,\%$$

Dies bedeutet, dass 78,6 % aller Sterbefälle durch Speiseröhrenkrebs unter den Exponierten auf Tabakkonsum zurückzuführen sind.

Im Bereich Public Health wird häufig das **bevölkerungsbezogene attributable Risiko** berechnet (s.a. Kap. 6.4.4). Es wird meistens als *Population Attributable Risk* (PAR) oder als *Population Attributable Fraction* (PAF) bezeichnet. Das bevölkerungsbezogene attributable Risiko gibt an, welcher Anteil der Fälle in der gesamten Bevölkerung auf die Exposition zurückgeht. Damit entspricht es auch dem Anteil der Fälle in der Bevölkerung, der vermeidbar wäre, wenn die Exposition beseitigt würde – vorausgesetzt, die Assoziation zwischen Exposition und Outcome ist wirklich ursächlich. Das bevölkerungsbezogene attributable Risiko wird nach folgender Formel berechnet:

$$\text{Bevölkerungsbez. attribut. Risiko} = \frac{\left(\begin{array}{c}\text{Kum. Inzidenz in}\\ \text{d. Gesamtbevöl.}\end{array}\right) - \left(\begin{array}{c}\text{Kum. Inzidenz unter den}\\ \text{Nichtexponierten}\end{array}\right)}{\text{Kum. Inzidenz in der Gesamtbevölkerung}} \; (\cdot 100)$$

Alternativ können Sie das bevölkerungsbezogene attributable Risiko mit Hilfe des *Relativen Risikos* (RR) und des Anteils der Exponierten in der Bevölkerung (p) ermitteln:

$$\text{Bevölkerungsbezogenes attributables Risiko} = \frac{p \cdot (RR - 1)}{p \cdot (RR - 1) + 1} \; (\cdot 100)$$

2.1.4 Validität und Reliabilität

Die Güte eines Messverfahrens wird durch seine Validität und Reliabilität bestimmt. Die **Validität** (Gültigkeit) bezeichnet das Ausmaß, in dem ein Messverfahren das misst, was es messen soll. Eine Messung gilt dann als valide, wenn ihr Ergebnis der Realität entspricht. Der Begriff der **Reliabilität** (Wiederholbarkeit) beschreibt das Ausmaß, in dem Messwerte replizierbar sind. Ein Messverfahren oder eine Messung gelten also dann als reliabel, wenn zu unterschiedlichen Messzeitpunkten oder bei Messwiederholungen gleiche Messwerte ermittelt werden. Eine Abbildung in Kap. 2.1

auf unserer Lehrbuch-Homepage zeigt schematisch am Beispiel der Bestimmung des Körpergewichts die Auswirkungen hoher bzw. geringer Validität und Reliabilität eines Messverfahrens auf das Verhältnis zwischen den gemessenen Werten und dem wahren (aber unbekannten) Wert.

In der Epidemiologie – etwa bei der Untersuchung von Ursache-Wirkungs-Beziehungen – wird darüber hinaus auch noch zwischen interner und externer Validität unterschieden. Diese Begriffe beziehen sich jedoch nicht auf die oben beschriebene Eigenschaft von Messungen und Messverfahren, sondern auf die Qualität der Studie und die Anwendbarkeit der Resultate. Eine hohe **interne Validität** liegt dann vor, wenn in einer Studie die beobachtete Ausprägung eines Outcomes allein auf die Exposition zurückzuführen ist und Alternativerklärungen für das Vorliegen oder die Höhe der gefundenen Effekte weitestgehend ausgeschlossen werden können. Dabei sinkt die interne Validität mit der steigenden Anzahl von plausiblen alternativen Erklärungen aufgrund von Fehlern und Verzerrungen. Bevor eine Untersuchung als intern valide bezeichnet werden kann, müssen deshalb Störgrößen als Ursache für einen beobachteten Zusammenhang ausgeschlossen werden (s. Kap. 2.1.8). Der Begriff der **externen Validität** bezeichnet die Anwendbarkeit oder Generalisierbarkeit der Studienergebnisse über die StudienteilnehmerInnen hinaus auf andere Populationen. Eine geringe externe Validität findet man bei unnatürlichen Untersuchungsbedingungen und bei geringer Repräsentativität der untersuchten Stichprobe. Eine hohe interne Validität ist Voraussetzung für eine hohe externe Validität.

2.1.5 Epidemiologische Studientypen

In der analytischen Epidemiologie werden Bevölkerungen oder Bevölkerungsgruppen im Hinblick auf die interessierenden Expositionen und Outcomes verglichen, um die oben erwähnten drei Fragen nach dem Vorhandensein, der Stärke und der Kausalität einer Assoziation (s. Kap. 2.1.1) zu beantworten. Analytisch-epidemiologische Studien im engeren Sinne sind beobachtende Studien. Hier wird die Exposition nicht von den Forschenden zugeteilt, sondern durch sie beobachtet. Die gängigsten Designs sind die *Querschnittstudie,* die *Kohortenstudie* und die *Fall-Kontroll-Studie*. Experimentelle Studien, bei denen ForscherInnen die Exposition zuteilen, werden in Kap. 2.1.6 besprochen.

Querschnittstudien

In einer Querschnittstudie werden alle Variablen, deren Assoziation untersucht werden soll, zum gleichen Zeitpunkt erhoben. Das erlaubt eine schnelle Durchführung, kann aber zu schwer interpretierbaren Ergebnissen führen. Da eine zeitliche Achse fehlt, bleibt oft unklar, welche Variable die Exposition ist und welche der Outcome. Wir stellen uns hierzu als Beispiel eine Studie vor, bei der in einem Kranken-

haus bei allen PatientInnen der Cholesterinspiegel gemessen wird. Gleichzeitig wird ermittelt, ob die/der PatientIn eine Krebserkrankung hat oder nicht. Es findet sich nun ein statistischer Zusammenhang zwischen einem niedrigen Cholesterinspiegel und einer Krebserkrankung. Nun wäre es falsch, auf der Basis solcher Querschnittdaten den Schluss zu ziehen, dass ein niedriger Cholesterinspiegel ein Risikofaktor für Krebserkrankungen ist. Da die Studie keine zeitlichen Informationen erhoben hat, könnte es ebenso gut sein, dass eine Krebserkrankung zu einem niedrigen Cholesterinspiegel führt – etwa, weil die Betroffenen den Appetit verlieren und nicht mehr genügend essen. Tatsächlich erscheint das sogar als die plausiblere Interpretation dieser Ergebnisse. Mit dem gewählten Querschnittdesign ist eine Klärung jedoch nicht möglich. Hierzu müsste eine Kohortenstudie durchgeführt werden. Untersuchungen, bei denen lediglich die Prävalenz *eines* Risikofaktors oder *einer* Erkrankung zu einem bestimmten Zeitpunkt gemessen wird, werden ebenfalls als Querschnittstudien oder auch als „Prävalenzstudien" bezeichnet. In diesem Fall handelt es sich jedoch um deskriptive – und nicht um analytische – Studien.

Kohortenstudien

Kohortenstudien werden auch als prospektive (d. h. in die Zukunft schauende) oder longitudinale (sich über einen Zeitraum erstreckende) Studien bezeichnet. Sie beginnen mit einer Gruppe von *Exponierten* und einer Gruppe von *Nichtexponierten*, die alle vom zu untersuchenden Outcome frei – also im Hinblick auf diesen Outcome gesund – sein müssen. Beide Gruppen werden über einen bestimmten Zeitraum nachverfolgt. In jeder der beiden Gruppen wird die Inzidenz des Outcomes berechnet (s. Abb. 2.1). Das ist in Kohortenstudien möglich, da sowohl die Zahl der Fälle als auch die Zahl der Personen unter Risiko bzw. deren Personenzeit (s. Kap. 2.1.2) bekannt sind. Als Assoziationsmaß dient das *Relative Risiko* (RR) oder die *Relative Rate* (s. Kap. 2.1.3).

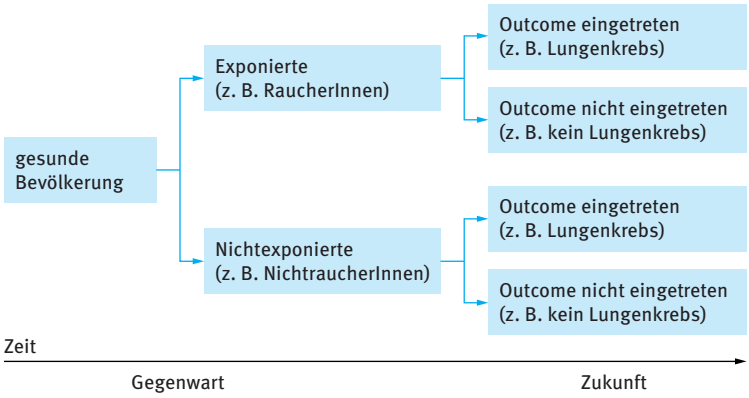

Abb. 2.1: Schema einer Kohortenstudie.

Kohortenstudien sind besonders geeignet, wenn
- mehrere Outcomes einer Exposition untersucht werden sollen (z. B. das Risiko von RaucherInnen im Vergleich zu NichtraucherInnen, verschiedene Krebsarten zu entwickeln)
- der Outcome häufig ist
- die Exposition selten ist
- der Expositionsstatus sich über die Zeit verändert (dies lässt sich durch Angabe der exponierten bzw. nicht exponierten Personenzeit berücksichtigen)

Viele chronische, nichtübertragbare Erkrankungen haben eine lange *Latenzzeit*, d. h. es existiert ein Zeitraum zwischen der Exposition und dem Auftreten des Outcomes. Daher dauern Kohortenstudien, die Ursachen solcher Krankheiten untersuchen, oft Jahre oder Jahrzehnte. Wird dagegen z. B. das Risiko der Entwicklung einer Fehlbildung als Folge einer Exposition während der Schwangerschaft erforscht, so muss die dazu durchgeführte Kohortenstudie nur auf die Dauer der Schwangerschaft angelegt sein.

Historische Kohortenstudien: Eine Kohortenstudie kann auch so angelegt werden, dass sie sich von einem Zeitpunkt in der Vergangenheit bis in die Gegenwart erstreckt. Das ist möglich, wenn für eine Personengruppe entsprechende Expositions- und Gesundheitsdaten aus der Vergangenheit vorliegen. Solche „historischen Kohortenstudien" werden z. B. in der Arbeitsepidemiologie durchgeführt, wenn in einem Unternehmen Unterlagen über Expositionen am Arbeitsplatz (z. B. zum Arbeiten mit einem Lösungsmittel) für die gesamte Beschäftigungsdauer vorliegen. Außerdem muss für alle Beschäftigten ermittelt werden, ob der Outcome (etwa eine Krebserkrankung) eingetreten ist. Die nicht exponierte Vergleichsgruppe sollte aus dem gleichen oder einem vergleichbaren, anderen Betrieb kommen, damit der sozioökonomische Status ähnlich ist. Eine Stichprobe aus der Allgemeinbevölkerung als Vergleichsgruppe kann zu einer Verzerrung führen, da ihr Gesundheitszustand im Schnitt schlechter sein kann als der einer arbeitenden Population (*Healthy-worker-Effekt*). Ein Beispiel für eine historische Kohortenstudie ist die *Swiss National Cohort*. Hierbei wurden die Daten, die Ende 1990 in einer Volkszählung (u. a. zum Bildungsstand der Schweizer Bevölkerung) erhoben wurden, mit der Statistik der Todesfälle der Jahre 1991 bis 2008 verlinkt.

Die „SAPALDIA Kohorte" und die „Nationale Kohorte": Im Normalfall wird vor dem Beginn einer Kohortenstudie entsprechend der Forschungsfrage genau festgelegt, welche Expositionen und welche Outcomes untersucht werden sollen. So untersuchte z. B. die *SAPALDIA Kohorte* (Swiss study on Air Pollution And Lung Disease in Adults) den Einfluss der Luftverschmutzung auf die Gesundheit der Atemwege und des Herz-Kreislauf-Systems. Es gibt jedoch auch Kohortenstudien, bei denen es keine vorformulierte Forschungsfrage gibt. In solchen Kohorten werden eine Vielzahl potenziell

interessanter Expositionen sowie möglicher Confounder (s. Kap. 2.1.8) gemessen. Gleichzeitig werden unterschiedliche Outcomes erfasst. Ein Beispiel dafür ist die *Nationale Kohorte* in Deutschland. Hier sollen 200.000 Menschen über einen Zeitraum von mindestens 10 bis 20 Jahren beobachtet werden. Ein solches Design erlaubt es den EpidemiologInnen, im Verlauf der Studie eine Vielzahl von Forschungsfragen zu generieren. Sie identifizieren dann jeweils exponierte und nicht exponierte Untergruppen in der Kohorte und ermitteln, ob der für die jeweilige Forschungsfrage interessierende Outcome eintritt.

Fall-Kontroll-Studien

Fall-Kontroll-Studien sind retrospektiv, also zeitlich gesehen „zurückschauend" (s. Abb. 2.2). Der Outcome ist bereits eingetreten, es wird nun der Expositionsstatus bei Personen mit Outcome (*Fälle*) und ohne Outcome (*Kontrollen*) erfragt und miteinander verglichen. Als Assoziationsmaß in Fall-Kontroll-Studien dient die *Odds Ratio* (OR), ein Näherungswert für das Relative Risiko (RR). Inzidenzraten lassen sich in Fall-Kontroll-Studien nicht berechnen, da nicht die gesamte Population, aus der die Fälle stammen, sondern nur ausgewählte Kontrollen in die Studie aufgenommen werden.
Fall-Kontroll-Studien sind besonders geeignet, wenn

- mehrere Expositionen untersucht werden sollen, die möglicherweise mit einem bestimmten Outcome assoziiert sind (z. B. verschiedene Expositionen, die das Risiko für Herz-Kreislauf-Erkrankungen erhöhen)
- der Outcome selten ist

Wollte man die Ursachen einer seltenen Krebserkrankung in einer Kohortenstudie untersuchen, so müsste man hunderttausende von Personen über einen längeren Zeitraum beobachten, um genügend inzidente (d. h. neu aufgetretene) Fälle zu

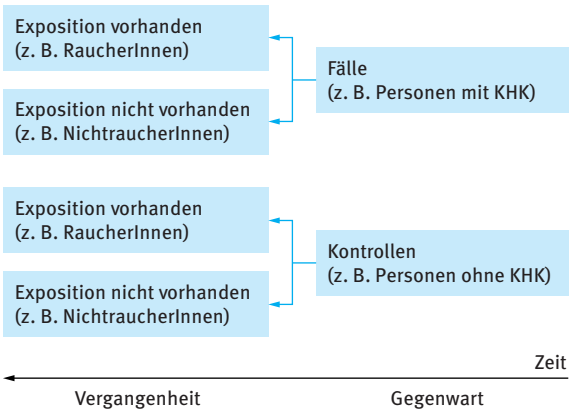

Abb. 2.2: Schema einer Fall-Kontroll-Studie.

finden. Einfacher und effizienter ist es, die Fälle, die über mehrere Jahre in einer großen Region aufgetreten sind, etwa mithilfe eines Krebsregisters zu identifizieren und dann zu befragen. Allerdings hängt die Qualität solcher Daten vom Erinnerungsvermögen der Fälle und der Kontrollen ab. Da die Rückfragezeiträume oft Jahre oder Jahrzehnte umfassen, kann es dabei zu Ungenauigkeiten kommen. Erinnern sich Fälle besser als Kontrollen – etwa, weil sie intensiv über mögliche Ursachen ihrer Krankheit nachdenken –, so kann das zu Verzerrungen führen (*Recall Bias*, s. Kap. 2.1.8).

Eine weitere Schwierigkeit bei Fall-Kontroll-Studien ist die Auswahl einer geeigneten Gruppe von Kontrollen. Idealerweise sollten sie als Zufallsstichprobe aus der gleichen Bevölkerung rekrutiert werden, aus der auch die Fälle stammen. Das ist mit einem hohen organisatorischen Aufwand verbunden, zudem ist die Teilnahmebereitschaft oft nur gering, was wiederum zu Verzerrungen führen kann (*Non-Response Bias*, s. Kap. 2.1.8). Eine mögliche Alternative sind Kontrollen aus einem Krankenhaus. Sie werden aus Abteilungen oder Stationen gezogen, in denen PatientInnen liegen, die das gesuchte Outcome nicht haben. In einer Studie zu oralen Kontrazeptiva („Pille“) und Brustkrebs könnten das beispielsweise Patientinnen einer orthopädischen Station sein. Es gibt aber immer wieder Belege dafür, dass sich KrankenhauspatientInnen hinsichtlich potenziell interessierender Expositionen wie Rauchen oder Alkoholkonsum von der Allgemeinbevölkerung unterscheiden. Sie sind also für diese nicht wirklich repräsentativ. Daraus kann wiederum eine Verzerrung aufgrund eines *Selektionsbias* (s. Kap. 2.1.8) resultieren.

Eine Fall-Kontroll-Studie kann nur dann durchgeführt werden, wenn die Exposition in der Bevölkerung nicht allzu selten ist, und wenn der Expositionsstatus von den Befragten erinnert werden kann. Letzteres ist nicht immer der Fall: Beim EHEC-Ausbruch 2011 (s. Kap. 9.2.3) in Deutschland waren Sprossen der Überträger der Erkrankung. Die ersten Fall-Kontroll-Studien im Rahmen der Ausbruchsuntersuchung konnten dies allerdings nicht nachweisen. Die EHEC-PatientInnen erinnerten sich nicht daran, dass sie Sprossen gegessen hatten. Die Sprossen waren meist Dekoration oder unauffällige Beigabe zu Salaten. Gut erinnert – und in der Folge fälschlich beschuldigt – wurden lediglich die optisch viel auffallenderen Gurken und Tomaten im Salat, die jedoch für die Übertragung gar nicht verantwortlich waren. Das Erfragen der Exposition und die Auswahl einer geeigneten Kontrollgruppe sind also besondere Herausforderungen bei Fall-Kontroll-Studien.

2.1.6 Klinische Studien

Klinische Studien werden mit PatientInnen oder gesunden ProbandInnen durchgeführt, um Medikamente, bestimmte Behandlungsformen oder andere medizinische Interventionen auf ihre Wirksamkeit und Sicherheit zu überprüfen. Klinische Studien sind meist experimentelle Studien. Anders als bei den in Kap. 2.1.5 vorgestellten beobachtenden Studientypen teilen ForscherInnen hier einer Studiengruppe eine

bestimmte Intervention zu. Die klinische Prüfung verläuft in vier Phasen (hier dargestellt am Beispiel einer Medikamentenstudie):

- *Phase I*
  Studien an einer kleinen Zahl gesunder ProbandInnen (20–80), um die Wirkungen eines Medikaments am potentiellen Wirkort (Organsystem) zu untersuchen. Im Blickpunkt stehen darüber hinaus die Pharmakokinetik, Verträglichkeit und Sicherheit der Substanz.
- *Phase II*
  Erprobung an einer größeren Zahl von PatientInnen (200–500), um erste Hinweise auf die Wirksamkeit des Präparats und die notwendige Dosierung zu erhalten.
- *Phase III*
  Studie der therapeutischen Wirksamkeit an einer großen Zahl von PatientInnen (einige hundert bis Tausende). Gelingt dieser Wirkungsnachweis, wird in der Regel die Marktzulassung für das Medikament beantragt. Bei den Phase-III-Studien handelt es sich um randomisierte, kontrollierte Studien (*Randomized Controlled Trials*, RCT).
- *Phase IV*
  Studien nach der Zulassung des Medikaments, um unter den behandelten PatientInnen mögliche unerwünschte Arzneimittelwirkungen erkennen zu können. Phase-IV-Studien sind *beobachtende Studien*, da hier die Zuteilung zu der mit einem bestimmten Medikament behandelten Gruppe in der Routineversorgung der Arztpraxis und nicht kontrolliert durch die ForscherInnen vorgenommen wird.

Randomisierte, kontrollierte Studien

In Phase-III-Studien wird die Wirksamkeit eines neuen Medikaments (*Verum*) mit der eines alten Medikaments oder der eines Scheinmedikaments (*Placebo*) verglichen. Die Gruppe, die das alte Medikament oder das Placebo erhält, wird auch als Kontrollgruppe oder Kontrollarm bezeichnet, die Verum-Gruppe auch als Interventionsgruppe oder Interventionsarm. Bei diesem Design kann es zu Verzerrungen kommen, wenn sich die PatientInnen der Interventions- und der Kontrollgruppe in Eigenschaften unterscheiden, die Einfluss auf das Behandlungsergebnis haben. Dies wäre z. B. der Fall, wenn ein Altersunterschied zwischen den Gruppen bestände und sich die Heilungsaussichten mit zunehmendem Alter verschlechtern (*Confounding*, s. Kap. 2.1.8). Auch wäre es denkbar, dass ÄrztInnen besonders schwer erkrankte PatientInnen bevorzugt in die Verum-Gruppe aufnehmen – in der Hoffnung, dass die PatientInnen dort besonders effektiv behandelt werden. In der Verum-Gruppe befänden sich nun mehr schwerer Erkrankte, die leichteren Fälle verblieben in der Kontrollgruppe. Als Folge davon würde die Wirksamkeit des neuen Medikaments unterschätzt.

Eine ungleiche Verteilung prognostischer Faktoren zwischen beiden Gruppen muss also vermieden werden. Um dieses Ziel zu erreichen, werden die StudienteilnehmerInnen zufallsgesteuert entweder der Interventionsgruppe oder der Kontrollgruppe zugewiesen. Ziel einer solchen *Randomisierung* ist es, Strukturgleichheit herzustellen: Mögliche Störfaktoren werden mit gleicher Wahrscheinlichkeit auf die Interventions- und die Kontrollgruppe verteilt. Bei größeren Studien mit mehreren hundert StudienteilnehmerInnen pro Arm kann davon ausgegangen werden, dass sich bekannte (und unbekannte) prognostische Faktoren gleichmäßig auf die Studiengruppen verteilen, womit eine verzerrende Wirkung verhindert wird.

Randomisierung: Von einer Randomisierung kann nur dann gesprochen werden, wenn die Zuteilung wirklich zufallsgesteuert erfolgt (z. B. durch computergenerierte Randomisierungslisten). Andere denkbare Verteilungsverfahren beinhalten die Möglichkeit einer Verzerrung. Wenn etwa alle PatientInnen, die montags in ein Krankenhaus aufgenommen werden, der Interventionsgruppe zugeteilt werden, während die Dienstags-PatientInnen in die Kontrollgruppe kommen, ist eine Strukturgleichheit beider Gruppen nicht gewährleistet. Es ist dann durchaus denkbar, dass montags (d. h. nach dem Wochenende) viele besonders schwere Fälle aufgenommen werden. Ein weiteres Problem ergibt sich aus der Vorhersehbarkeit der Zuordnung: Die für die Aufnahme der PatientInnen verantwortlichen Personen können die Zuordnung beeinflussen, indem sie z. B. diejenigen PatientInnen montags in die Studie aufnehmen, die ihrer Ansicht nach stärker von einer Therapie mit dem neuen Medikament profitieren. Die Zuordnung muss deshalb auch bei Verwendung von Randomiserungslisten immer verdeckt durch eine außenstehende Stelle vorgenommen werden, die unabhängig vom direkt in die Studie involvierten Personal ist (*Concealment of Allocation*).

Ist die Zahl der StudienteilnehmerInnen gering, wählt man in der Regel eine *stratifizierte Randomisierung*. Hierdurch können wichtige Merkmale der PatientInnen gleichmäßig auf die Interventions- und die Kontrollgruppe verteilt werden, ohne das Prinzip der Randomisierung zu unterlaufen. Dies kann z. B. dann angebracht sein, wenn vermutet wird, dass ein Impfstoff bei Männern und Frauen unterschiedlich wirkt. Die PatientInnen werden dabei zunächst in Geschlechtergruppen aufgeteilt (stratifiziert). Anschließend wird die Randomisierung jeweils innerhalb dieser Geschlechtergruppen vorgenommen. Mit der *Blockrandomisierung* erreicht man, dass jeweils gleich viele PatientInnen den vorhandenen Studiengruppen zugeordnet werden. Dabei wird für jeden PatientInnen-Block (z. B. für jeweils 8 Personen) zufallsgesteuert festgelegt, in welcher Reihenfolge sie der Interventions- bzw. der Kontrollgruppe zugeteilt werden. Hierbei muss das zuvor definierte Verhältnis eingehalten werden (z. B. vier PatientInnen in jeder der beiden Gruppen). So ist sichergestellt, dass die entstehenden Gruppen zufallsverteilt gleich groß sind.

Verblindung: Zu einer Verzerrung der Studienergebnisse kann es auch dann kommen, wenn z. B. ÄrztInnen das Behandlungsergebnis unbewusst als besonders positiv beurteilen, weil sie wissen, dass der/die untersuchte PatientIn das Verum (und nicht das Placebo) erhalten hat. UntersucherInnen können also durch die Kenntnis der Studienhypothese und der Gruppenzugehörigkeit in ihrer Beurteilung des Behandlungsergebnisses beeinflusst werden. Um eine solche Verzerrung zu vermeiden, wird eine so genannte Verblindung durchgeführt (*Blinding*). Hierbei wissen weder PatientInnen noch UntersucherInnen, ob das Verum oder ein Placebo gegeben wurde. Voraussetzung dafür ist, dass sich Verum- und Placebopräparat optisch gleichen. In der Regel werden heute Phase-III-Studien als randomisierte, kontrollierte Studien (RCT) durchgeführt, bei denen sowohl PatientInnen als auch untersuchende ÄrztInnen verblindet wurden (*Doppelblind-Studie*). In einigen Studien werden auch diejenigen Personen verblindet, die die Daten bearbeiten. Für sie ist zunächst nicht erkennbar, welche Gruppe die Verum- und welche die Placebo-Gruppe ist. Die Informationen hierzu werden von der Studienleitung unter Verschluss gehalten. Erst nach der Durchführung der Analyse wird offengelegt, welche Gruppe welches Präparat erhalten hat. Eine Verblindung aller Beteiligten ist nicht immer möglich (z. B. beim Vergleich von invasiven mit konservativen Therapien). Von der Verblindung zu unterscheiden ist die oben beschriebene *verdeckte Zuordnung* (Concealment of Allocation) bei der Randomisierung. Sie hilft Ungleichheiten der Studiengruppen bei der Randomisierung zu vermeiden und ist immer machbar.

Studienablauf: Bei der Planung einer RCT ist in einem ersten Schritt zu klären, welche PatientInnen die Voraussetzungen für die Teilnahme an der Studie erfüllen. Dazu müssen klare Ein- und Ausschlusskriterien definiert sein. Anschließend werden die potentiellen StudienteilnehmerInnen über das Ziel der Studie, die Randomisierung und die Verblindung aufgeklärt. Wenn sie ihr Einverständnis zur Teilnahme an der Studie gegeben haben (*Informed Consent*), werden sie in die Studie aufgenommen. Erst danach findet die Randomisierung statt. Die PatientInnen erhalten dann entweder das Verum- oder das Placebopräparat. Abb. 2.3 zeigt schematisch den Ablauf einer solchen Studie.

Die Auswertung der Daten erfolgt ähnlich wie bei einer Kohortenstudie. Es können z. B. die kumulativen Inzidenzen für Outcomes oder für das Auftreten von Nebenwirkungen in beiden Gruppen verglichen werden. Das Ergebnis wird dann als relatives Risiko angegeben. Auch in diesem Stadium einer klinischen Studie können sich Verzerrungen ergeben, etwa wenn PatientInnen nicht in der Behandlungsgruppe verbleiben, der sie durch die Randomisierung zugewiesen wurden. So wäre es denkbar, dass ein behandelnder Arzt eine Patientin bei einem besonders schweren Verlauf und fehlendem Hinweis auf Besserung der anderen Gruppe zuweist – in der Hoffnung, es könne sich dabei um die Verum-Gruppe mit einem wirksamen Medikament handeln. Auf diese Weise würde sich die Zahl der PatientInnen mit schwererem Verlauf in einer der beiden Gruppen erhöhen, was das Ergebnis verzerren würde. Um dies zu

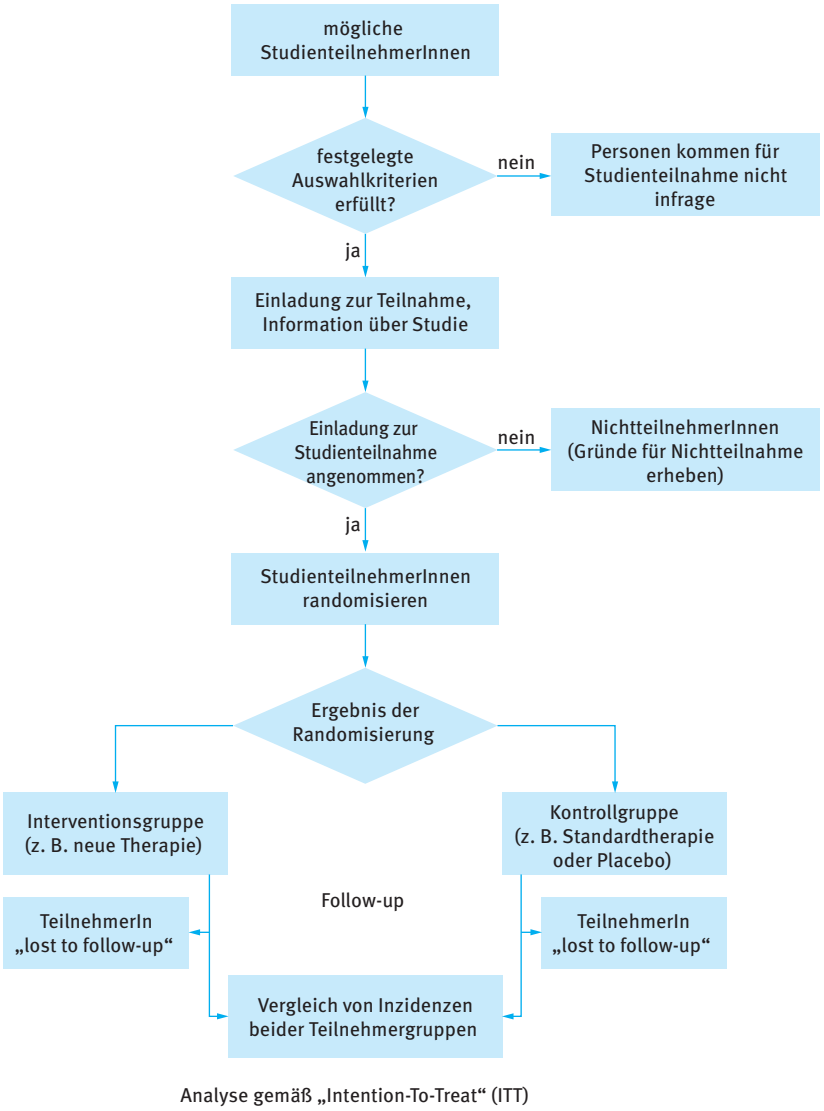

Abb. 2.3: Schematischer Ablauf einer randomisierten, kontrollierten Studie (RCT).

vermeiden, werden alle PatientInnen entsprechend ihrer ursprünglichen Zuordnung ausgewertet, unabhängig davon, welche Behandlung sie tatsächlich erhalten haben. Diese Analyse aufgrund der beabsichtigten Therapie wird mit *Intention-to-Treat* (ITT) bezeichnet und sollte in klinischen Studien Standard sein. Probleme entstehen auch dann, wenn PatientInnen ihre Teilnahme an der Studie aufkündigen oder der Kontakt zu den StudienteilnehmerInnen abreißt (*Loss to follow up*, LTFU).

Fallzahlkalkulation (Sample Size Calculation): Vor dem Beginn einer klinischen Studie muss zunächst die erforderliche Größe der Studie festgelegt werden. Mit Unterstützung eines Statistikers wird hierzu eine Fallzahlkalkulation vorgenommen. Zunächst ist zu überlegen, wie groß der Unterschied zwischen der Verum- und Kontrollgruppe mindestens sein muss, um klinisch relevant zu sein. Anschließend wird das *Signifikanzniveau* (s. Kap. 2.3.8) festgelegt – in der Regel entscheidet man sich für ein Signifikanzniveau von 5 %. Das bedeutet, dass die Wahrscheinlichkeit, diesen Unterschied (oder einen größeren Unterschied) zu beobachten, wenn in Wirklichkeit *kein* Unterschied besteht, kleiner als 5 % sein soll. Weiter muss die so genannte *Power* der Studie festgelegt werden. Die Power gibt die Wahrscheinlichkeit an, dass die Studie den Unterschied, wenn er wirklich besteht, als statistisch signifikant aufdeckt. Meist entscheidet man sich für eine Power von 80 %.

Wird eine klinische Studie mit einer zu kleinen Fallzahl durchgeführt, so sind die Ergebnisse statistisch nicht aussagekräftig, denn die Power der Studie reicht nicht aus, um einen möglicherweise bestehenden Unterschied aufzuzeigen. In diesem Fall wird das 95 %-Vertrauensintervall (siehe Kap. 2.3.6) des Relativen Risikos oder der Relativen Rate die 1 mit einschließen, d. h. die Ergebnisse für das Verumpräparat sind im Vergleich zum Placebo oder dem alten Medikament sowohl mit einer besseren als auch mit einer schlechteren Wirkung vereinbar. Die Studie hat ihren Zweck dann nicht erfüllt. Leider sind viele klinische Studien in der Tat zu klein, um klinisch relevante Unterschiede aufzuzeigen oder auszuschließen. In dieser Situation können *Meta-Analysen* mehrerer Studien sinnvoll sein (siehe Kap. 2.1.7).

Relative und absolute Risikoreduktion: Bei der Analyse der Daten einer randomisierten, kontrollierten Studie werden zunächst die *absoluten Risiken* oder kumulative Inzidenzraten (s. Kap. 2.1.2) in den beiden Therapiegruppen berechnet. Die *relative Risikoreduktion* (RRR) gibt an, um welchen Prozentsatz der Einsatz der Verum-Therapie das Ergebnis gegenüber der Kontrollgruppe verändert. Die relative Risikoreduktion berechnet sich aus der kumulativen Inzidenz in den beiden Studienarmen.

Durch die Multiplikation mit 100 wird die relative Risikoreduktion in Prozent angeben.

$$\text{RRR} = \frac{\text{Kumulative Inzidenz}_{\text{Kontrollgruppe}} - \text{Kumulative Inzidenz}_{\text{Interventionsgruppe}}}{\text{Kumulative Inzidenz}_{\text{Kontrollgruppe}}} \cdot 100$$

Die relative Risikoreduktion ist wenig aussagekräftig, wenn der untersuchte Outcome sehr selten ist. Eine Senkung einer sehr geringen Wahrscheinlichkeit um einen bestimmten Prozentsatz ist möglicherweise klinisch auch nicht relevant. Ein besseres Maß ist hier die *absolute Risikoreduktion* (ARR). Sie errechnet sich aus der Differenz der Risiken bzw. Inzidenzraten in beiden Therapiegruppen.

$$\text{ARR} = \text{Kumulative Inzidenz}_{\text{Kontrollgruppe}} - \text{Kumulative Inzidenz}_{\text{Interventionsgruppe}}$$

Ist die ARR größer als 0, so wirkt die Verum-Therapie besser als die Therapie, die in der Kontrollgruppe eingesetzt wurde (altes Medikament oder Placebo). Wirkt die neue Therapie jedoch nicht besser oder sogar schlechter als die Kontrollmedikation, ist die absolute Risikoreduktion gleich oder kleiner als 0. Da die ARR statt einer prozentualen Veränderung die tatsächliche Inzidenz angibt, ist die klinische Relevanz eines so angegebenen Ergebnisses leichter zu beurteilen.

Number Needed to Treat/Number Needed to Harm: Ein gebräuchliches Maß, um die Wirksamkeit eines neuen Medikaments zu beschreiben, ist die *Number Needed to Treat* (NNT). Sie gibt an, wie viele PatientInnen mit dem neuen anstatt dem alten Medikament behandelt werden müssen, um einen zusätzlichen Behandlungserfolg zu erzielen. Die NNT wird auch als „Anzahl der erforderlichen Behandlungen" bezeichnet. Sie berechnet sich folgendermaßen:

$$\text{NNT} = \frac{1}{\text{ARR}} = \frac{1}{\text{Kumulative Inzidenz}_{\text{Kontrollgruppe}} - \text{Kumulative Inzidenz}_{\text{Interventionsgruppe}}}$$

Eine weitere, wichtige Messgröße ist die *Number Needed to Harm* (NNH). Sie gibt die Zahl der PatientInnen an, die mit der neuen Therapie behandelt werden müssen, bis im Vergleich zur Kontrollgruppe bei einem/r PatientIn eine zusätzliche unerwünschte Wirkung auftritt. Die NNH zeigt damit, wie häufig unerwünschte, durch das Verum-Präparat hervorgerufene Wirkungen (= Nebenwirkungen) sind. Ihre Berechnung erfolgt entsprechend dem Vorgehen bei der NNT.

Die Berechnung von RRR, ARR und NNT werden im Folgenden an einem Beispiel illustriert. In einer Studie zur Wirksamkeit eines neuen Cholesterinsenkers erhielten 2.221 PatientInnen das neue Medikament und 2.223 ein Placebo. Im Laufe des Follow-ups von durchschnittlich 5,4 Jahren verstarben 182 PatientInnen in der Interventionsgruppe (Medikament) und 256 PatientInnen in der Kontrollgruppe (Placebo). Das Sterberisiko in der Interventionsgruppe betrug daher 182/2.221 = 0,08 (oder 8 %), das in der Kontrollgruppe 256/2.223 = 0,12 (oder 12 %). Die relative Risikoreduktion (RRR), die absolute Risikoreduktion (ARR) und die Number Needed to Treat (NNT) lassen sich nun entsprechend den oben angegebenen Formeln wie folgt berechnen:

RRR: (0,12–0,08) / 0,12 · 100 = 33,3 %

ARR: (0,12–0,08) = 0,04 (oder 4 %)

NNT: 1/0,04 = 25.

Der Einsatz des neuen Cholesterinsenkers reduzierte damit die Sterblichkeit gegenüber der Kontrollgruppe um ca. 33 %, d. h. in der Interventionsgruppe starben ein Drittel weniger PatientInnen als in der Kontrollgruppe. Absolut betrachtet konnte

das Medikament die Sterblichkeit jedoch nur um 4 % senken. Um einen Todesfall zu verhindern, müssen nun 25 PatientInnen über einen Zeitraum von durchschnittlich 5,4 Jahren mit dem neuen Medikament behandelt werden.

Ethische Aspekte

Die ethischen Prinzipien, die in einer klinischen Studie berücksichtigt werden müssen, sind in der *Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects* (s. Internet-Ressourcen) aufgelistet. Diese Deklaration fordert, das genaue Versuchsprotokoll einer Ethikkommission vorzulegen. Die Ethikkommission prüft, ob die geplante Studie ethisch vertretbar und zulässig ist. Hierfür wird u. a. beurteilt, ob die möglichen Nachteile, die mit einer Studienteilnahme verbunden sind, in einem vertretbaren Verhältnis zu den möglichen Vorteilen der Studie stehen (Nutzen-Risiko-Verhältnis). Weiterhin wird geprüft, ob Patienteninformation und Einwilligungserklärung umfassend und verständlich sind. Ethikkommissionen sind unabhängige Gremien und bestehen meist aus ÄrztInnen, Pflegefachkräften, StatistikerInnen, JuristInnen und MedizinethikerInnen oder TheologInnen.

Internationale Richtlinien der *guten klinischen Praxis* (s. Internetressourcen) definieren Standards, die die Durchführung einer klinischen Studie regeln. Die Einhaltung dieser Standards gewährleistet die Qualität der Studien und erleichtert Vergleiche zwischen verschiedenen Studien. Die Berichterstattung und Dokumentation zu randomisierten klinischen Studien sollte auf der Basis der *CONSORT*-Richtlinien (*Consolidated Standards of Reporting Trials Statement*, s. Internet-Ressourcen) erfolgen. Zu einer solchen ordnungsgemäßen Berichterstattung gehört, dass die Ergebnisse *aller* randomisierten, kontrollierten Studien publik gemacht werden – auch wenn sie gezeigt haben, dass ein neues Präparat nicht wirksamer ist als ein altes Präparat oder ein Placebo. Eine solche Offenlegung aller Studienergebnisse ist besonders wichtig für die Durchführung von Meta-Analysen (s. Kap. 2.1.7), weil auf diese Weise eine Verzerrung der Resultate durch einen *Publikationsbias* (eine Form von *Selektionsbias*, s. Kap. 2.1.8) verhindert werden kann.

Heute sind viele wissenschaftliche Zeitschriften nur dann bereit, eine randomisierte, kontrollierte Studie zu veröffentlichen, wenn die Studie zuvor in ein spezielles Register aufgenommen wurde (z. B. *Deutsches Register Klinischer Studien* oder *EU Clinical Trials Register*). Die von der Weltgesundheitsorganisation (WHO) unterstützten und koordinierten Register sollen verhindern, dass negative Ergebnisse bei klinischen Studien unter Verschluss bleiben.

Randomisierte, kontrollierte Studien bei komplexen Interventionen

Randomisierte, kontrollierte Studien kommen auch außerhalb der klinischen Forschung zum Einsatz, etwa um die Wirksamkeit von Maßnahmen der Verhaltensprävention (z. B. der Raucherentwöhnung, s. Kap. 4.4), von Screening-Programmen (s. Kap. 4.5.3) oder Versorgungsmodellen (z. B. *Managed Care*) zu evaluieren. In der

Tat wird zunehmend gefordert, auch Public-Health-Interventionen in randomisierten, kontrollierten Studien zu erproben und damit eine stärkere Evidenzbasierung von Public Health zu schaffen (*Evidence-based Public Health*, EbPH).

Auch bei diesen randomisierten, kontrollierten Studien teilen ForscherInnen einer Studiengruppe eine bestimmte Intervention zu, und auch hier gibt es eine Kontrollgruppe. Die Interventionen sind im Gegensatz zu klinischen Medikamentenstudien jedoch oft komplexer und Placebo-Kontrollen in der Regel nicht möglich. Dies erschwert die Verblindung der StudienteilnehmerInnen oder macht sie unmöglich. Ein weiteres Problem ist die *Kontamination* des Interventionseffektes, die dadurch entsteht, dass sich StudienteilnehmerInnen aus Interventions- und Kontrollgruppe austauschen. Die Interventionsmaßnahmen gelangen so in die Kontrollgruppe. Um dies zu verhindern, werden oft nicht einzelne Personen, sondern Arztpraxen, Pflegestationen, Dörfer oder Versorgungsregionen randomisiert. Ein weiterer Grund für eine solche *Cluster-Randomisierung* liegt darin, dass sich Interventionen oft nicht auf der individuellen Ebene umsetzen lassen.

2.1.7 Systematische Übersichten und Meta-Analysen

Angesichts der großen Anzahl publizierter Einzelstudien sind gute Übersichten, die Auskunft über die vorhandene Evidenzlage geben, für die Entscheidungsfindung in der klinischen Medizin und in der Gesundheitspolitik besonders wichtig. Wir unterscheiden in diesem Zusammenhang zwischen narrativen und systematischen Übersichtsarbeiten einerseits und Meta-Analysen andererseits.

Narrative Übersichten fassen die Ergebnisse wichtiger Studien informell zusammen und kommentieren sie. Die Auswahl und Beurteilung der Studien ist subjektiv und nicht immer umfassend. Die Schlussfolgerungen, die die AutorInnen ziehen, spiegeln nicht zuletzt auch ihre persönliche Meinung wider. Gerade aus diesem Grund haben narrative Übersichtsarbeiten jedoch auch weiterhin ihren Platz in der Literatur.

Im Unterschied zu narrativen Übersichtsarbeiten zeichnen sich **systematische Übersichtsarbeiten** durch eine klar definierte Fragestellung und ein reproduzierbares, wissenschaftliches Vorgehen aus. Es handelt sich hierbei um Studien über Studien. Nicht Personen werden untersucht, sondern alle Studien, die die zuvor festgelegten Einschlusskriterien erfüllen. In systematischen Übersichtsarbeiten wird die methodologische Qualität der in die Arbeit aufgenommenen Studien nach festgelegten Kriterien beurteilt, die Ergebnisse werden standardisiert erfasst. Ein- und Ausschlusskriterien, die Kriterien für die Studienqualität, Outcomes und Interventionen werden ebenso wie die Literatursuche ausführlich beschrieben. Auch mögliche Verzerrungen durch *Publikationsbias* (s. Kap. 2.1.6 und Kap. 2.1.8) oder eine mangelhafte Qualität der Studien werden diskutiert. Auf diese Weise werden die Schlussfolgerungen der AutorInnen leicht nachvollziehbar und reproduzierbar. Die *Cochrane Collaboration* ist ein internationales Netzwerk von WissenschaftlerInnen, ÄrztInnen

und PatientInnen, das qualitativ hochstehende systematische Übersichtsarbeiten verfasst. Ihr Ziel ist es, dadurch evidenzbasiertes Handeln in der medizinischen Versorgung und in Public Health zu fördern. Ihre Übersichtsarbeiten werden über die *Cochrane Library* zugänglich gemacht.

Systematische Reviews enthalten oft eine oder mehrere **Meta-Analysen**. In Meta-Analysen werden die Ergebnisse der verschiedenen Studien statistisch zusammengefasst, um präzisere und allgemeingültigere Angaben über die Wirksamkeit von Interventionen zu erhalten. Einzelne Studien sind oft nicht groß genug, um kleinere, aber klinisch bedeutende Unterschiede sicher zu erfassen. Zudem sind die Ergebnisse von Einzelstudien oft nur auf eine relativ eng umschriebene Population anwendbar. Im Zentrum der Meta-Analyse steht die Berechnung eines *Summationswertes* mit einem

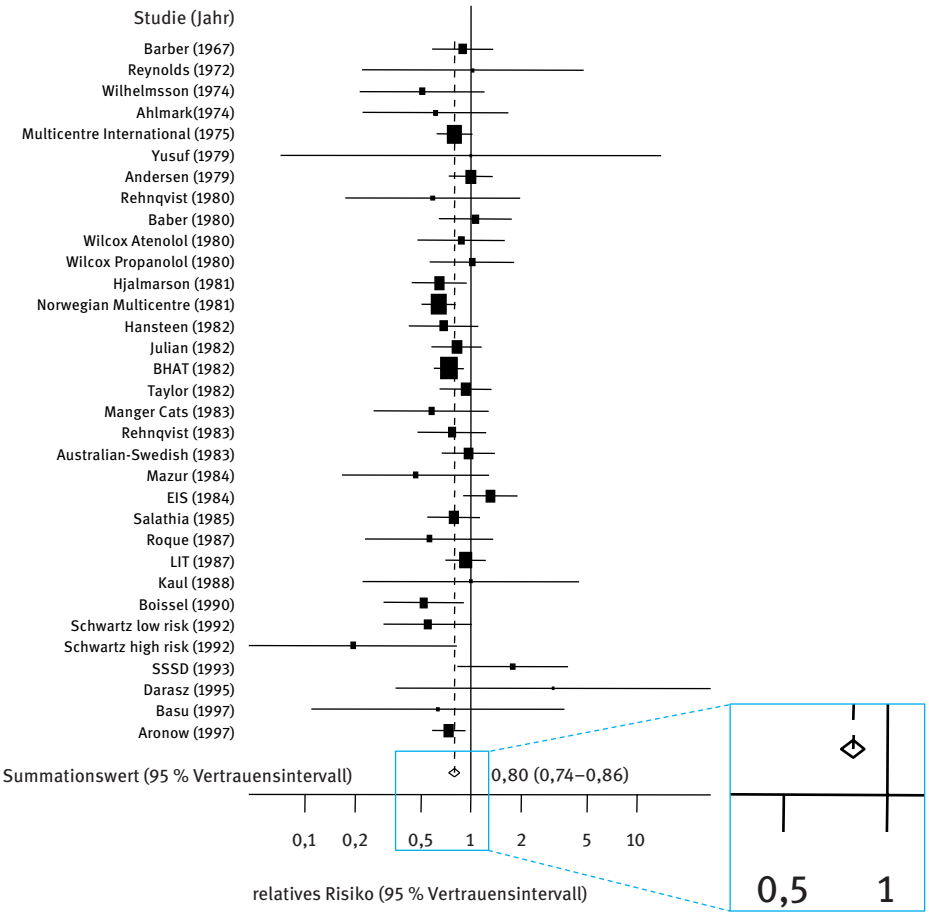

Abb. 2.4: Meta-Analyse von randomisierten, klinischen Studien zur sekundärpräventiven Wirksamkeit einer Therapie mit Betablockern bei PatientInnen nach Myokardinfarkt.

Vertrauensintervall. Beim Summationswert handelt es sich um einen gewichteten Mittelwert, der große Studien stärker gewichtet als kleinere Studien. Der Summationswert wird zusammen mit den Ergebnissen der einzelnen Studien in einem *Forest Plot* dargestellt.

Abb. 2.4 zeigt den Forest Plot einer Meta-Analyse von randomisierten klinischen Studien, die die sekundärpräventive Wirkung von verschiedenen Betablockern auf die Mortalität bei Personen nach einem Herzinfarkt untersuchten. Die schwarzen Rechtecke repräsentieren die gefundene Wirkung auf das Sterblichkeitsrisiko in den einzelnen Studien. Die Flächen der Rechtecke zeigen dabei das Gewicht der Studien in der Meta-Analyse, die horizontalen Linien die 95 %-Vertrauensintervalle. Die vertikale Linie stellt das relative Risiko von 1 dar. Hier gibt es also keinen Unterschied in der Wirkung zwischen Betablocker- und Kontrollgruppen. In Studien, die links von dieser Linie liegen, war die Sterblichkeit in der Betablocker-Gruppe niedriger als in der Kontrollgruppe, während auf der rechten Seite der Linie die Kontrollgruppen besser abschnitten. Die gestrichelte Linie zeigt den Summationswert aus der Meta-Analyse an, das Diamant-Zeichen unterhalb der aufgelisteten Einzelstudien veranschaulicht das 95 %-Vertrauensintervall des Summationswertes (s. vergrößerter Ausschnitt in Abb. 2.5). *Achtung:* In Forest Plots werden logarithmische Skalen verwendet, sodass 0,5 und 2 den gleichen Abstand zu 1 aufweisen und die Vertrauensintervalle damit symmetrisch sind!

In den meisten 95 %-Vertrauensintervallen unserer Abbildung ist die 1 enthalten. Die Mehrzahl der Studien zeigt somit, dass Betablocker keine statistisch signifikante sekundärpräventive Wirkung auf die Mortalität bei Personen nach einem Herzinfarkt haben ($p \geq 0{,}05$). Durch die Meta-Analyse wird hingegen deutlich, dass die Einnahme von Betablockern in dieser Situation zu einer Reduktion der Sterblichkeit um 20 % führt. Das errechnete enge Vertrauensintervall schließt die 1 hier nicht ein (relatives Risiko = 0,80; 95 %-Vertrauensintervall 0,74–0,86). Anhand der gestrichelten Linie ist eine visuelle Beurteilung der Variabilität bzw. Heterogenität der Ergebnisse der Studien möglich. In unserem Beispiel ist diese recht klein. Die Ergebnisse der Studien liegen nahe beieinander, die 95 %-Vertrauensintervalle schließen den Summationswert aus der Meta-Analyse mit ein. Es handelt sich somit um eine homogene Situation. Dies legt die Schlussfolgerung nahe, dass der Wirkung der verschiedenen Betablocker ein Klasseneffekt[12] zugrunde liegt, der in unterschiedlichen Patientenpopulationen zum Tragen kommt und deshalb eine hohe externe Validität (Generalisierbarkeit) aufweist.

Meta-Analysen können die Wirksamkeit von Medikamenten und anderen Interventionen oft früher und überzeugender nachweisen als einzelne kleinere Studien. In diesem Fall wäre der Nutzen der Betablocker bereits Anfang der 1980er Jahre nachweisbar gewesen. Die Methode sollte jedoch keinesfalls unkritisch angewendet

12 *Klasseneffekt*: Ein Effekt, der bei einer Wirkstoff-Klasse und nicht nur bei einem einzelnen Medikament auftritt.

werden. Grundsätzlich kann die Qualität einer Meta-Analyse nicht besser sein als diejenige der Einzelstudien (*Garbage in – Garbage out*[13]). Auch wird der Heterogenität der Ergebnisse oft nicht genügend Beachtung geschenkt. Wenn stark voneinander abweichende Resultate vorliegen, ist eine statistische Kombination dieser Ergebnisse selten sinnvoll. Schließlich kann ein *Publikationsbias* (s. Kap. 2.1.8) in den ausgewählten Studien die Ergebnisse von Meta-Analysen verzerren.

2.1.8 Mögliche Fehlerquellen in epidemiologischen Untersuchungen

Bevor beurteilt werden kann, ob eine ermittelte *Assoziation* zwischen einer Exposition und einem Outcome kausal ist, d. h. ob hier eine Ursache-Wirkungs-Beziehung besteht, muss zunächst untersucht werden, inwiefern mögliche Fehler und Verzerrungen den wahren Zusammenhang zwischen beiden verschleiern.

Zufällige Fehler

Ein zufälliger Fehler liegt dann vor, wenn der Wert einer gemessenen Variablen in einer Stichprobe rein zufallsbedingt um den tatsächlichen (wahren) Wert der Variablen in der Grundgesamtheit streut. Eine häufige Quelle von zufälligen Fehlern sind ungenaue Messungen. Ist die Stichprobe groß genug, fallen Messungenauigkeiten in der Regel nicht ins Gewicht. Die einzelnen Messwerte schwanken dann zwar um den wahren Wert und die *Reliabilität* der Messung ist gering, denn bei jeder Messung wird ein anderer Wert ermittelt. Im Durchschnitt nähern sich die einzelnen Messwerte aber dem wahren Wert an, die *Validität* der Messung ist daher hoch (s. Kap. 2.1.4).

Um mit ausreichender Sicherheit auf die Grundgesamtheit schließen zu können, ist auch bei genauen Messungen eine ausreichend große Stichprobe nötig. Wenn die Stichprobe zu klein ist, können trotz genauer Messungen zufällige Fehler in Form von so genannten *Stichprobenfehlern* auftreten. Dies lässt sich leicht mit Hilfe eines Würfels veranschaulichen. Wird nur zehnmal gewürfelt, kann es passieren, dass manche Augenzahlen oft, andere vielleicht gar nicht gewürfelt werden, obwohl bei einem „fairen" Würfel alle sechs Augenzahlen die gleiche Wahrscheinlichkeit haben, geworfen zu werden (nämlich ein Sechstel). Stichprobenfehler nehmen mit zunehmender Stichprobengröße ab. Wie groß die Stichprobe letztendlich sein muss, um den zufälligen Fehler auf einem akzeptablen und vor Studienbeginn definierten Niveau zu halten, ist von unterschiedlichen Faktoren abhängig und Gegenstand der *Fallzahlplanung* (s. Kap. 2.1.6).

13 *Garbage in – Garbage out* (engl.): Wo man Müll hineinsteckt, kommt auch Müll heraus.

Systematische Fehler

Von einem systematischen Fehler oder **Bias** ist in der Epidemiologie dann die Rede, wenn der Fehler nicht zufällig, sondern immer in der gleichen Weise auftritt. Ein Beispiel hierfür ist eine ungeeichte Waage, die das Gewicht einer Person immer um 5 Kilogramm zu hoch angibt. Eine solche Messung ist *reliabel*, denn bei jeder Messung wird der gleiche Wert gemessen. Ihre Validität ist jedoch gering, da auch mit zunehmender Anzahl von Messungen der korrekte (wahre) Wert nicht ermittelt werden kann (vgl. Kap. 2.1.4).

Selektionsbias: Ein Selektionsbias bezeichnet einen systematischen Auswahlfehler bei der Rekrutierung von StudienteilnehmerInnen oder bei deren Verbleib in einer Studie. Ein Beispiel für einen häufigen Auswahlfehler ist ein sogenannter *Non-Response-Bias*. Er kann auftreten, wenn sich die TeilnehmerInnen einer Studie von den Personen unterscheiden, die eine Studienteilnahme verweigern. Ein Grund hierfür ist, dass Eigenschaften, die mit der Teilnahmeverweigerung zusammenhängen, oft auch mit der Exposition und dem Outcome assoziiert sind. So konsumieren Teilnahmeverweigerer (so genannte *Non-Responder*) im Vergleich zu TeilnehmerInnen oft mehr Alkohol und Tabak. Um den Grad der Verzerrung abzuschätzen, die durch einen Non-Response-Bias entsteht, müssen daher auch grundlegende Informationen über Non-Responder eingeholt werden. Eine spezielle Variante des Non-Response-Bias kann in der Beobachtungs-Phase (*Follow-up*) von Kohorten- und randomisierten kontrollierten Studien auftreten. Dies ist der Fall, wenn sich die Studienausfälle in der Gruppe der Exponierten und Nichtexponierten (bzw. der Interventions- und Kontrollgruppe) in Faktoren unterscheiden, die auch mit der Exposition oder dem Outcome in Zusammenhang stehen. Um diesen so genannten *Loss-to-follow-up-Bias* zu vermeiden, müssen StudienteilnehmerInnen in Längsschnittstudien intensiv nachverfolgt werden (vgl. Abb. 2.3).

Ein Selektionsbias kann nicht nur StudienteilnehmerInnen, sondern auch wissenschaftliche Ergebnisse betreffen. WissenschaftlerInnen und angesehene wissenschaftliche Publikationsorgane neigen nämlich dazu, eher Ergebnisse zu veröffentlichen, die Zusammenhänge zwischen Expositionen und Outcomes aufzeigen. Studien, die (wider Erwarten) keine Assoziationen erkennen lassen, werden daher oft gar nicht oder in Zeitschriften veröffentlicht, die wenig Beachtung finden. Für Meta-Analysen stellt dieser so genannte *Publikationsbias* ein Problem dar, denn er kann zu einer Überschätzung des tatsächlichen Zusammenhangs zwischen Expositionen und Outcomes führen (s. Kap. 2.1.7).

Informationsbias: Ein Informationsbias (auch *Missklassifikation* genannt) bezeichnet einen systematischen Messfehler, der dazu führt, dass Personen im Hinblick auf Exposition, Outcome und weitere Einflussvariablen (*Kovariaten*) falsch klassifiziert werden. Exponierte werden z. B. fälschlicherweise den Nichtexponierten zugeordnet oder Kranke den Gesunden.

Die möglichen Folgen eines Informationsbias werden im Folgenden an der in Kap. 2.1.3 beschriebenen Fall-Kontroll-Studie zur Assoziation von Neuroleptikaeinnahme und venöser Thromboembolie (VTE) illustriert. Die AutorInnen ermittelten in ihrer Untersuchung eine Odds Ratio von 1,6. Wir nehmen nun an, dass sowohl bei den Fällen (VTE) als auch den Kontrollen (keine VTE) 5 % aller Personen, die keine Neuroleptika einnahmen, dies bei der Befragung nicht korrekt angaben und damit fälschlicherweise der Gruppe der Exponierten zugeordnet wurden. Eine solche Missklassifikation, die unabhängig vom Outcome- oder Expositionsstatus ist, wird als *nicht-differenzielle Missklassifikation* bezeichnet. Die Vier-Felder-Tafel hierzu ist in Tab. 2.3 dargestellt.

Tab. 2.3: Vier-Felder-Tafeln zum Zusammenhang von Neuroleptikaeinnahme und venöser Thromboembolie (VTE; Zahlen nach Parker et al. 2010, s. Tab. 2.2) mit einer hypothetischen fünfprozentigen *nicht-differenziellen Missklassifikation* (oben) und einer hypothetischen *differenziellen Missklassifikation* (unten).

(1) Bei der nicht-differenziellen Missklassifikation wurden 5 % aller Personen, die keine Neuroleptika einnahmen, fälschlicherweise der Gruppe der Exponierten zugeordnet.
(2) Die differenzielle Missklassifikation beträgt bei den erkrankten Personen 10 % (d.h. 10 % aller Personen, die keine Neuroleptika einnahmen, wurden fälschlicherweise der Gruppe der Exponierten zugeordnet) und bei den nicht erkrankten Personen 5 %.

Nicht-differenzielle Missklassifikation

		Outcome (VTE)		
		Ja	*Nein*	*Summe*
Exposition (Neuroleptikaeinnahme)	*Ja*	3.296 a	8.989 b	12.285 a+b
	Nein	22.236 c	80.502 d	102.738 c+d
	Summe	25.532 a+c	89.491 b+d	115.023 a+b+c+d

Odds Ratio = (3.296 · 80.502) / (8.989 · 22.236) = 1,33

Differenzielle Missklassifikation

		Outcome (VTE)		
		Ja	*Nein*	*Summe*
Exposition (Neuroleptikaeinnahme)	*Ja*	4.467 a	8.989 b	13.456 a+b
	Nein	21.065 c	80.502 d	101.567 c+d
	Summe	25.532 a+c	89.491 b+d	115.023 a+b+c+d

Odds Ratio = (4.467 · 80.502) / (8.989 · 21.065) = 1,90

Die Odds Ratio ist bei der nicht-differenziellen Missklassifikation mit (3.296 · 80.502) / (8.989 · 22.236) = 1,33 geringer als zuvor. Wie in diesem Beispiel führt eine nicht-differenzielle Missklassifikation fast immer zu einer Unterschätzung des Zusammenhangs von Exposition und Outcome.

Dagegen können Missklassifikationen, die vom Outcome- oder Expositionsstatus abhängig sind (so genannte *differenzielle Missklassifikationen*), sowohl zu einer Unter- als auch zu einer Überschätzung der Assoziation von Exposition und Outcome führen. Zu einer differenziellen Missklassifikation kann es z. B. infolge eines unterschiedlichen Erinnerungsvermögens bei Fällen und Kontrollen kommen (*Recall-Bias*) – ein Problem, das häufig in Fall-Kontroll-Studien auftritt. Ein solcher Recall-Bias könnte beispielsweise in der genannten Fall-Kontroll-Studie dazu führen, dass die Missklassifikation bei den erkrankten Personen mit 10 % höher ist als bei den nicht erkrankten Personen, wo sie nur 5 % beträgt (Tab. 2.3). In diesem hypothetischen Fall einer differenziellen Missklassifikation beträgt die Odds Ratio 1,9. Die Assoziation wird daher im Vergleich zu den Originaldaten überschätzt.

Confounding

Der Begriff Confounding bezeichnet eine Verzerrung, die durch den Einfluss einer oder mehrerer weiterer Einflussvariablen (so genannter *Confounder* oder *Störgrößen*) entsteht, sodass die Assoziation zwischen Exposition und Outcome über- oder unterschätzt wird.

Confounding liegt dann vor, wenn diese Einflussvariablen sowohl mit dem Outcome als auch mit der Exposition assoziiert sind. Ein Beispiel für Confounding ist die Assoziation von gelben Fingern (*Exposition*) und Lungenkrebs (*Outcome*). Menschen, die viel rauchen, haben bedingt durch das Kondensat des Zigarettenrauchs oft gelbe Finger. Rauchen ist aber auch ein bekannter Risikofaktor für Lungenkrebs, und zwar unabhängig davon, ob die RaucherInnen gelbe Finger haben oder nicht. Das Rauchen ist also im Hinblick auf den Zusammenhang zwischen gelben Fingern und Lungenkrebs ein Confounder. Wird dieser Confounder nicht berücksichtigt, ergibt sich (fälschlicherweise) ein Zusammenhang zwischen dem Vorhandensein von gelben Finger und Lungenkrebs. Abb. 2.5 veranschaulicht das mit Hilfe eines so genannten *Confounding-Dreiecks*.

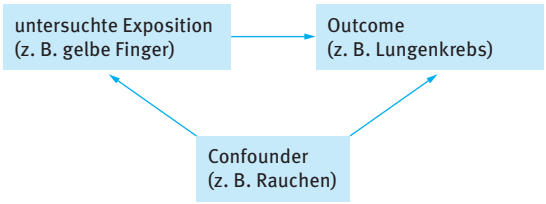

Abb. 2.5: Schematische Darstellung von Confounding am Beispiel der Assoziation von gelben Fingern und Lungenkrebs; nach: Davey Smith G, Phillips AN. Confounding in epidemiological studies: why "independent" effects may not be all they seem. British Medical Journal 1992; 305: 757–759.

Häufige Confounder in epidemiologischen Untersuchungen sind das *Alter* und der *sozioökonomische Status*. Es gibt verschiedene Möglichkeiten, potenzielle Confounder bei der Durchführung und Analyse von Studien zu berücksichtigen.

- So kann man bereits im Studiendesign vorsehen, nur Personen in die Studie einzuschließen, die im Hinblick auf mögliche Confounder homogen sind. Man würde dann z. B. nur Personen ähnlichen Alters oder mit einem ähnlichen sozioökonomischen Status in die Studie aufnehmen (Einschränkung der Studienbevölkerung).
- In Fall-Kontroll-Studien besteht darüber hinaus die Möglichkeit, für potenzielle Confounder zu matchen (*Matching*). Hierzu wird jedem Fall eine Kontrolle zugeordnet, die ihm in der Ausprägung eines oder mehrerer Confounder ähnelt.
- In randomisierten kontrollierten Studien (s. Kap. 2.1.6) werden Personen nach dem Zufallsprinzip der Interventions- und Kontrollgruppe zugeordnet. Auch potentielle Confounder werden dadurch gleichmäßig zwischen beiden Gruppen verteilt, sodass *Strukturgleichheit* entsteht.
- Wenn eine epidemiologische Studie bereits durchgeführt wurde, ist es möglich, Confounding durch eine *stratifizierte Analyse* zu kontrollieren. Dazu wird die Assoziation zwischen Exposition und Outcome getrennt für die einzelnen Ausprägungen (*Strata;* von lat. *Stratum* = Schicht) des vermeintlichen Confounders ermittelt. Confounding liegt dann vor, wenn die ermittelte Assoziation größer oder kleiner ist, als bei der unstratifizierten Analyse errechnet wurde.
- Werden mehrere Confounder vermutet, können diese durch *multivariate Verfahren* kontrolliert werden.

Die beiden letzten Möglichkeiten setzen voraus, dass potenzielle Confounder in der Studie erhoben wurden.

Variablen, die Zwischenstufen auf dem kausalen Pfad von Exposition und Outcome darstellen – so genannte *Intermediärvariablen* – sind keine Confounder. Ein Beispiel hierfür ist „geringes Geburtsgewicht" bei Kindern von Frauen, die während der Schwangerschaft geraucht haben. Die *Exposition,* das Rauchen während der Schwangerschaft, führt hier über ein geringes Geburtsgewicht (*Intermediärvariable*) zu einer erhöhten Sterblichkeit der Säuglinge in der ersten Lebenswoche (*Outcome*).

Effektmodifikation

Die Stärke der Assoziation zwischen Exposition und Outcome kann sich je nach Ausprägung einer dritten Variablen – dem so genannten *Effektmodifikator* – unterscheiden. Dies wird als **Effektmodifikation** oder *Interaktion* bezeichnet. Tab. 2.4 zeigt die Ergebnisse einer Meta-Analyse, in der untersucht wurde, ob es einen Unterschied im Zusammenhang zwischen der *Exposition* Hepatitis-B-Virus-Infektion und dem *Outcome* Leberkrebs bei Rauchern und Nichtrauchern gibt. Die Vergleichsgruppe

besteht bei allen angegebenen Relativen Risiken aus NichtraucherInnen, die nicht mit dem Hepatitis-B-Virus (HBV) infiziert, also HBV-negativ, sind.

Die Untersuchung zeigt, dass NichtraucherInnen, die mit dem Hepatitis-B-Virus (HBV) infiziert sind, im Vergleich zu HBV-negativen NichtraucherInnen ein 15,8-fach höheres Risiko haben, ein Karzinom der Leber zu entwickeln. Das Risiko von HBV-positiven RaucherInnen, ein Leberkarzinom zu bekommen, ist jedoch mit einem RR von 21,6 noch deutlich höher. Beide Faktoren (HBV-Infektion und Rauchen) wirken also zusammen und verstärken sich gegenseitig. Anders als Confounding und Bias ist eine Effektmodifikation keine Verzerrung. Es ist jedoch ebenso wichtig, sie zu identifizieren und in der Auswertung zu berücksichtigen, da sonst wichtige Zusammenhänge unerkannt bleiben. Effektmodifikationen können wie Confounding mittels stratifizierter oder multivariater Analyse aufgedeckt werden.

Tab. 2.4: Relatives Risiko für die Entwicklung von Leberkrebs in Abhängigkeit von einer vorherigen Hepatitis-B-Virus-Infektion (Exposition) und von der Frage, ob die betroffene Person Raucher ist (Effektmodifikator).

		Effektmodifikator (Rauchen)	
		Nein	Ja
Exposition (Hepatitis-B-Virus-Infektion)	Nein	1,0 (Referenz)	1,9
	Ja	15,8	21,6

Quelle der Originaldaten: Chuang SC, Lee YC, Hashibe M, Dai M, Zheng T, Boffetta P. Interaction between cigarette smoking and hepatitis B and C virus infection on the risk of liver cancer: a meta-analysis. Cancer Epidemiology, Biomarkers and Prevention 2010; 19: 1261–1268.

Nur Assoziation oder auch Ursache?

Um Empfehlungen für Public-Health-Maßnahmen aussprechen zu können, reicht es nicht aus, die Assoziation zwischen Exposition und Outcome fehler- und verzerrungsfrei zu bestimmen. Zusätzlich muss ermittelt werden, ob ein Ursache-Wirkungs-Zusammenhang zwischen Exposition und Outcome besteht, die Assoziation also auch kausal ist. Einen statistischen Test auf Kausalität gibt es nicht. Daher müssen alle Hinweise für und gegen einen kausalen Zusammenhang sorgfältig beurteilt werden. *Sir Austin Bradford Hill* benannte schon 1965 verschiedene Kriterien, die es erleichtern, eine Assoziation auf ihre Kausalität hin zu beurteilen. Sie werden als **Bradford-Hill-Kriterien** bezeichnet:

- *Stärke der Beziehung*: Die Wahrscheinlichkeit einer kausalen Beziehung zwischen Exposition und Outcome nimmt mit der Stärke der Assoziation zu.
- *Konsistenz der Beziehung:* Wird eine Assoziation in mehreren unterschiedlichen Bevölkerungen (in verschiedenen Ländern, bei Personen unterschiedli-

chen Alters etc.) und mittels unterschiedlicher Studientypen festgestellt, ist eine kausale Beziehung wahrscheinlicher, als wenn dies nicht der Fall ist.
- *Zeitliche Sequenz:* Bei einer kausalen Beziehung muss die Ursache der Wirkung zwingend vorausgehen. Querschnittstudien sind daher weniger gut geeignet als Kohorten- und experimentelle Studien, auf eine eventuell vorhandene Kausalität zu schließen, da sie Expositionen und Assoziationen gleichzeitig erheben.
- *Spezifität des Effekts:* Dieses Kriterium fordert, dass eine Exposition (z. B. das Masernvirus) nur mit einem einzigen Outcome (hier: einer Masernerkrankung) assoziiert ist. Für die Untersuchung von Expositionen wie Rauchen oder Alkoholkonsum, die das Risiko für viele verschiedene Outcomes erhöhen, ist es weniger hilfreich.
- *Dosis-Wirkungs-Beziehung:* Steigt mit zunehmender „Menge" der Exposition (z. B.: 1–10 Zigaretten/Tag, 11–20 Zigaretten/Tag, 21–30 Zigaretten/Tag) das Risiko, dass das Outcome eintritt, ist eine Kausalität wahrscheinlicher als ohne Dosis-Wirkungs-Beziehung.
- *Biologische Plausibilität und Kohärenz:* Eine Ursache-Wirkungs-Beziehung zwischen Exposition und Outcome sollte biologisch plausibel sein und nicht im Widerspruch zu den Ergebnissen anderer Fachgebiete stehen.
- *Experimentelle Evidenz:* Kann man mittels einer randomisierten kontrollierten Studie zeigen, dass das Risiko für einen Outcome sinkt, sobald die Exposition beseitigt ist, liegt sehr wahrscheinlich eine Ursache-Wirkungs-Beziehung vor.

Die Bradford-Hill-Kriterien können die Beurteilung einer Ursache-Wirkungs-Beziehung zwischen Exposition und Outcome unterstützen, sie sollten jedoch nicht als Checkliste missverstanden werden. Nur selten wird eine Ursache-Wirkungs-Beziehung so deutlich wie bei der Assoziation zwischen Rauchen und Lungenkrebs. Hier werden mit Ausnahme der *Spezifität des Effekts* alle oben genannten Kriterien erfüllt.

2.1.9 Evidenzbasierte Medizin und Public Health

Das Konzept der *Evidence-based Medicine* (EbM) wurde in den 1980er Jahren an der McMaster-Universität in Hamilton, Kanada, von einer Gruppe um den Internisten und Epidemiologen **David Sackett** entwickelt. Gemeint ist hiermit eine Medizin, die von Ärztinnen und Ärzten nicht nur klinische Fertigkeiten und Erfahrung verlangt, sondern explizit auch Wissen aus der aktuellen patientenorientierten Forschung sowie Kenntnisse darüber, wie dieses Wissen zu interpretieren und anzuwenden ist. EbM zielt somit darauf ab, die Prinzipien und Methoden der klinischen Epidemiologie und der Gesundheitsökonomie (s. a. Kap. 2.1.6 und 2.5) in die klinische Praxis zu integrieren. Nach *Sackett* muss das Ziel dabei sein, einen *„gewissenhaften, ausdrücklichen und umsichtigen Gebrauch der aktuell besten wissenschaftlichen Daten für*

Entscheidungen in der Versorgung eines individuellen Patienten“ sicherzustellen. In die Entscheidungsfindung miteinbezogen werden sollen dabei neben der klinischen Situation selbstverständlich auch die Präferenzen der PatientInnen (Abb. 2.6).

Eine wichtige Fertigkeit im Rahmen der EbM, die erst in den letzten Jahre Eingang in die Curricula fand, ist das rasche Auffinden von relevanten Einzelstudien, systematischen Übersichten und evidenzbasierten Zusammenfassungen in medizinischen Datenbanken (z. B. in *PubMed*, *Cochrane Library*, *Clinical Evidence*, *UpToDate* etc.). Je nach Fragestellung sind hierbei randomisierte kontrollierte Studien zur Wirksamkeit von Medikamenten (s. a. Kap. 2.1.6), diagnostische Studien zu den Eigenschaften eines Tests (s. a. Kap. 2.3.7, Kap. 4.5), Kohortenstudien zur Prognose einer Erkrankung (s. a. Kap. 2.1.5), Kosten-Nutzen-Studien (s. a. Kap. 2.5) oder auch qualitative Studien etwa zur Compliance oder zur Zufriedenheit der PatientInnen mit einer bestimmten Therapiemethode (s. a. Kap. 2.4.5) gefragt.

In einem nächsten Schritt erfolgt dann die kritische Beurteilung der gefundenen Evidenz (*Critical Appraisal*). Dabei wird die interne Validität der Studien systematisch erfasst. Wichtige Kriterien sind hierbei Selektionsbias, Informationsbias und Confounding. Anschließend werden die Resultate und mögliche Zufallsfehler gewürdigt (s. a. Kap. 2.1.8) und die Anwendbarkeit im Rahmen des gegebenen klinischen Problems überprüft (externe Validität, s. a. Kap. 2.1.4). Nützliche Checklisten für dieses Vorgehen bei verschiedenen Studientypen sind zum Beispiel auf der Webseite des britischen *Critical Appraisal Skills Programme* abrufbar (CASP, siehe Internet-Ressourcen). Aufgrund der so gewonnenen Evidenzlage werden schließlich Empfehlungen in Form von Leitlinien erarbeitet. Tabelle 4.3 in Kap. 4.4.2 zeigt eine Kategorisierung von Empfehlungen für präventivmedizinische Maßnahmen aufgrund der Qualität der vor-

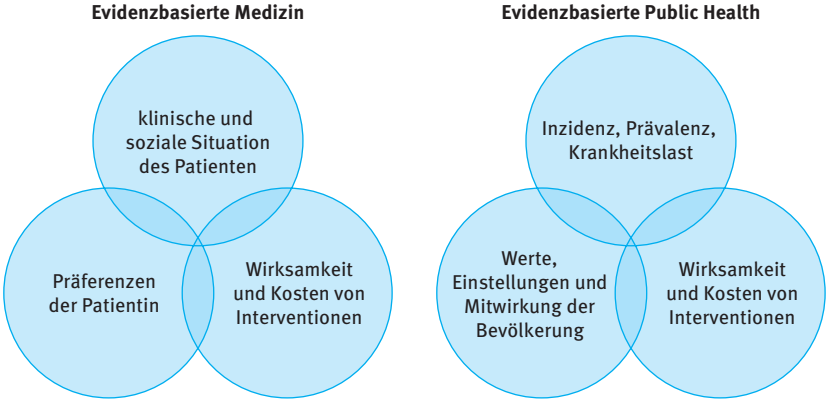

Abb. 2.6: Grundlagen für evidenzbasierte Entscheidungen in der klinischen Praxis und in Public Health.

handenen Evidenz. Das *GRADE-System* (*Grading of Recommendations, Assessment, Development and Evaluation*) vereinheitlichte die verschiedenen Klassifikationssysteme zur Bewertung der Evidenz und Formulierung von Empfehlungen. GRADE hat in den letzten Jahren stark an Bedeutung gewonnen und wird von der Weltgesundheitsorganisation (WHO), der *Cochrane Collaboration* und anderen Organisationen verwendet (s. Internet-Ressourcen).

Die Prinzipien der EbM sind sinngemäß auch auf die Pflege (*Evidence-based Nursing*), die Gesundheitsversorgung (*Evidence-based Health Care*) und die öffentliche Gesundheit (*Evidence-based Public Health*) anwendbar. Bei der evidenzbasierten Public Health (EbPH) verschiebt sich die Handlungsebene allerdings vom Patienten zum Gesundheitssystem bzw. zur Bevölkerung. So wird im Modell der EbPH die Erfassung der klinischen und sozialen Situation des einzelnen Patienten durch die Krankheitslast in der Bevölkerung (*Burden of Disease*, s. a. Kap. 10.1.2) ersetzt. Die Werte und Einstellungen sowie die Bereitschaft zur Mitwirkung in der Bevölkerung treten hier an die Stelle der individuellen Präferenzen (Abb. 2.6).

2.2 Demografie

Marcel Zwahlen, Nicole Steck, Matthias Egger

Die Frage „Wie viele sind wir?“ bewegt Regierungen bereits seit dem Altertum. Sie bildet die Grundlage der *Demografie* [von *démos* (gr.): Volk und *grafé* (gr.): Schrift, Beschreibung], die sich mit verschiedenen Merkmalen von Bevölkerungen beschäftigt. Dabei interessieren neben der Gesamtgröße der Bevölkerung, ihrer altersmäßigen Zusammensetzung und ihrer geografischen Verteilung auch die sozialen und Umweltfaktoren, die hier für Veränderungen verantwortlich sind. Die Daten zur fortlaufenden Beschreibung der Bevölkerung stammen mehrheitlich aus staatlichen Quellen, v. a. aus Volkszählungen, dem Geburten- und Sterberegister sowie repräsentativen Stichproben-Erhebungen.

2.2.1 Die Bevölkerung

Das Lukasevangelium berichtet über eine Anordnung des römischen Kaisers Augustus, nach der sich alle Bewohner des Reiches für eine Volkszählung in ihre Herkunftsorte zu begeben hatten. Maria und Josef reisten daraufhin nach Bethlehem, wo Jesus geboren wurde. Die Registrierung der Bevölkerung gab den Verantwortlichen in Rom einen Überblick über die Anzahl ihrer Steuerbürger. Die einfachste Information über eine Bevölkerung bezieht sich also auf die Zahl der Personen, die sich in einer geografisch definierten Region an einem bestimmten Datum befinden. Doch wenn Sie beispielsweise am 15. Juli die Anzahl an Personen zählen, die sich auf Mallorca befin-

den, erhalten Sie möglicherweise nicht die Zahl, die in *Wikipedia* unter „Bevölkerung von Mallorca“ aufgeführt wird (876.147 Einwohner für das Jahr 2012). Denn der Monat Juli ist Ferienzeit, und Sie werden Personen zählen, die nicht auf Mallorca wohnen, sondern sich nur vorübergehend dort aufhalten. Um die Bevölkerungszahl in einer Region zu ermitteln, zählt man daher diejenigen Personen, die an einem bestimmten Datum in dieser Region langfristig wohnhaft oder angemeldet sind. Dies setzt ein funktionierendes An- und Abmeldesystem beim Einwohneramt voraus. Auch mit einem solchen System kann es aber Personen geben, die zwar dort wohnen, aber nicht angemeldet sind oder Personen, die dort gemeldet sind, sich aber tatsächlich meistens anderswo aufhalten.

Die in der *Schweiz* wohnhafte Bevölkerung hat im Zeitraum von 1900 bis 2014 kontinuierlich von 3,3 auf 8,1 Mio. Einwohner zugenommen. Dies entspricht einer Zunahme um rund 150 %. Allein seit 1960 ist die Zahl der Einwohner um mehr als 50 % angewachsen. Aufgrund ihrer wechselhaften Geschichte lassen sich die Bevölkerungszahlen in Österreich erst ab 1947, in Deutschland erst ab 1956 gut darstellen. Die Einwohnerzahlen für *Deutschland* beziehen sich bis 1989 auf die beiden deutschen Staaten (vormalige Bundesrepublik Deutschland und Deutsche Demokratische Republik). Insgesamt hat die deutsche Bevölkerung im Zeitraum von 1960 bis 2013 von 72,6 auf 81,5 Mio. Einwohner zugenommen, das entspricht einer Zunahme um 13 % (Abb. 2.7). Die Größe der Bevölkerung *Österreichs* ist vergleichbar mit der der Schweiz, die Bevölkerungsentwicklung dort jedoch eher mit der Deutschlands. Die Zahl der Einwohner nahm von 6,9 Mio. im Jahr 1950 auf 8,5 Mio. im Jahr 2014 zu, was einer Zunahme um 23 % entspricht.

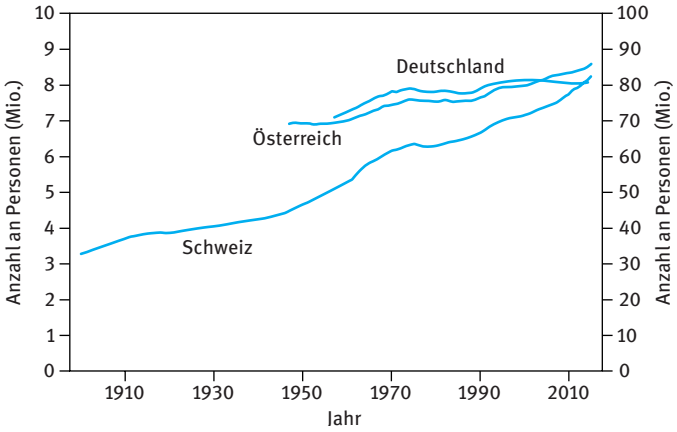

Abb. 2.7: Bevölkerungsentwicklung in Deutschland, Österreich und der Schweiz. Die Skala links bezieht sich auf die Bevölkerungszahlen von Österreich und der Schweiz, die rechte Skala auf die entsprechenden Zahlen für Deutschland.

Die Bevölkerungszahl kann durch natürliche Bevölkerungsbewegungen (Geburten und Todesfälle) oder durch räumliche Bevölkerungsbewegungen, also durch Ein- und Auswanderungen, zu- bzw. abnehmen.

Geburtenüberschuss und Geburtendefizit

Je nachdem, ob in einem bestimmten Gebiet mehr Geburten oder mehr Todesfälle aufgetreten sind, bezeichnet man die in einem gegebenen Kalenderjahr errechnete Differenz zwischen der Anzahl an Geburten und der Anzahl an Sterbefällen als *Geburtenüberschuss* oder *Geburtendefizit*. Die Differenz wird dann auf die aktuelle Bevölkerungszahl bezogen, d. h. durch die Gesamtzahl der Einwohner dividiert und pro 1.000 Einwohner angegeben. Somit ist der Wert nun zwischen verschiedenen Ländern und Gebieten vergleichbar. Deutschland meldet seit 1991 ein Geburtendefizit: Pro 1.000 Einwohner werden seit 1998 ein bis zwei Geburten weniger registriert als Todesfälle. In der Schweiz sank der Geburtenüberschuss von knapp 10 pro 1.000 Einwohner im Jahr 1960 auf 2,3 pro 1.000 Einwohner im Jahr 1980. Seither schwankt er zwischen 2 und 3 pro 1.000 Einwohner. In Österreich ist die Geburtenbilanz seit rund 20 Jahren mit kleinen Schwankungen in den positiven und den negativen Bereich praktisch ausgeglichen.

Migrationssaldo

Die Differenz zwischen der Zahl der Einwanderungen (*Immigration*) und der Zahl der Auswanderungen (*Emigration*) über Gebietsgrenzen hinweg wird als Migrations- oder Wanderungssaldo bezeichnet. Auch diese Größe wird meist für ein gegebenes Kalenderjahr berechnet und als absolute Zahl angegeben oder auf die Gesamtbevölkerung bezogen (pro 1.000 Einwohner). Der Wanderungssaldo kann kurzfristig stark schwanken. So führte in der Schweiz die durch die Erdölkrise von 1973 ausgelöste Rezession zu einem negativen Migrationssaldo in den Jahren 1975 bis 1977. Arbeitskräfte, die vor 1970 aus anderen Ländern in die Schweiz eingewandert waren, wanderten nun wieder vermehrt aus. Ausschlaggebend für die kontinuierliche Zunahme der Schweizer Bevölkerung von 6,6 Mio. Einwohnern im Jahr 1990 auf 7,95 Mio. Einwohner im Jahr 2012 ist primär ein positiver Migrationssaldo. In Österreich führte der seit Mitte der 1980er Jahre positive Migrationssaldo ebenfalls zu einer stetigen Zunahme der Einwohnerzahl. Auch Deutschland verzeichnete in den Jahren nach der Wiedervereinigung als Folge eines positiven Wanderungssaldos eine Bevölkerungszunahme, obwohl ein Geburtendefizit vorlag. Noch unklar ist, wie sich die derzeit hohe Zahl an Asylsuchenden und temporären Flüchtlingen längerfristig auf die Trends im Wanderungssaldo auswirken.

2.2.2 Entwicklung der Altersstruktur der Bevölkerung

Neben der Gesamtzahl der Einwohner liefert die Zusammensetzung der Bevölkerung nach Alter und Geschlecht wichtige Informationen. Abb. 2.8 zeigt die Altersstruktur der weiblichen und männlichen Schweizer Bevölkerung im Jahre 1900 sowie im Jahr 2015, untergliedert in Altersgruppen von jeweils einem Jahr. Im Jahr 1900 glich die Altersstruktur einer Pyramide, hier nahm die Anzahl an Personen mit ansteigendem Alter ab. Die Altersstruktur im Jahr 2015 lässt sich dagegen eher mit einer Urne oder einem Pilz vergleichen.

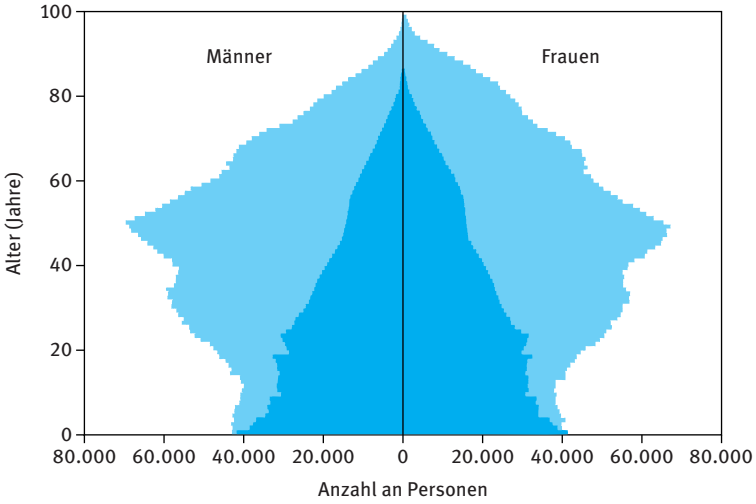

Abb. 2.8: Altersstruktur der Bevölkerung in der Schweiz in den Jahren 1900 (dunkelgrün) und 2015 (hellgrün).

Bei der Betrachtung der Altersstruktur in Deutschland im Jahr 1960 fallen besonders die Einbuchtungen im Alter von 41/42 Jahren und 15 Jahren auf (Abb. 2.9). Sie markieren das Ende des ersten bzw. zweiten Weltkrieges. Selbst in der Altersstruktur für das Jahr 2014 ist die Einbuchtung am Ende des zweiten Weltkrieges noch immer sichtbar, nun – 70 Jahre später – bei den etwa 70-Jährigen. Darüber hinaus wird deutlich, dass Frauen eine höhere Lebenserwartung als Männer haben. Dies führt dazu, dass es mehr Frauen als Männer in der Altersgruppe der über 80-Jährigen gibt (s. Kap. 2.2.4).

Sowohl in Deutschland und Österreich (s. Abb. 2.10) als auch in der Schweiz gibt es aktuell deutlich mehr Menschen im Alter von 45 bis 65 Jahren als im Alter zwischen 25 bis 45 Jahren. In den nächsten 15 bis 25 Jahren wird ein Großteil dieser 45- bis 65-Jährigen in Rente gehen. Sofern es in absehbarer Zukunft nicht mehr Geburten oder eine starke Einwanderung von jungen Menschen gibt, wird der Anteil der über 65-Jährigen in den nächsten beiden Dekaden in allen drei Ländern gegenüber 2010 stark

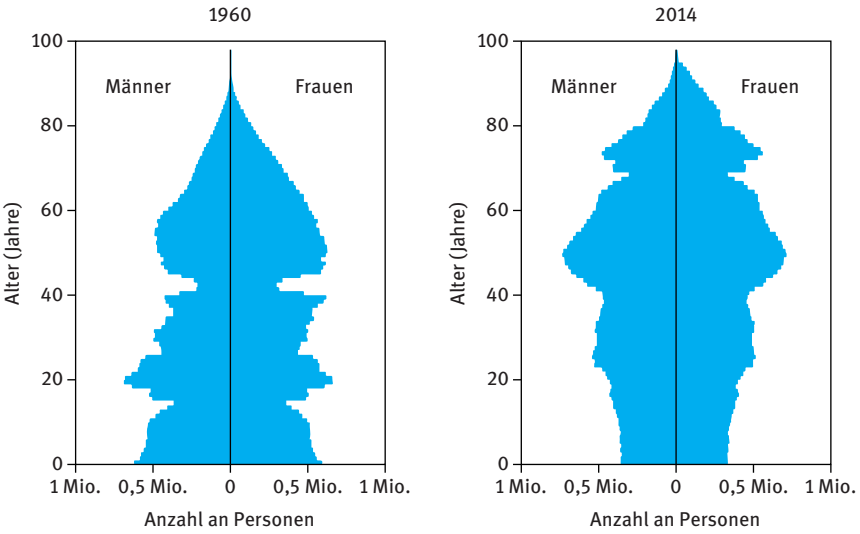

Abb. 2.9: Altersstruktur der Bevölkerung in Deutschland (West- und Ost-Deutschland zusammengezählt) in den Jahren 1960 und 2014.

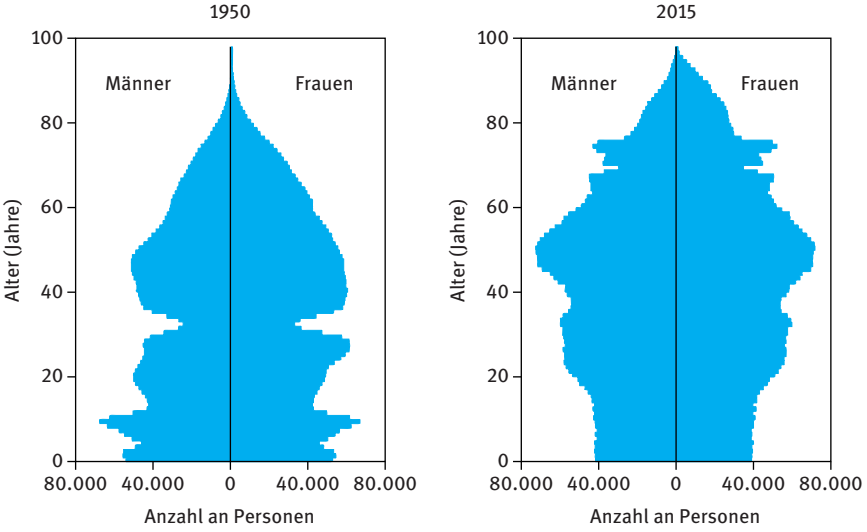

Abb. 2.10: Altersstruktur der Bevölkerung in Österreich in den Jahren 1950 und 2015.

zunehmen: in der Schweiz von derzeit 17 % auf fast 25 %, in Österreich von knapp 18 % auf gut 22 % und in Deutschland von 21 % auf fast 29 % (s. Tab. 2.5). Auch hier ist noch nicht klar, wie sich die in den Jahren 2015/2016 nach Deutschland eingereisten Asylsuchenden und temporären Flüchtlinge auf diese Trends auswirken werden. Da viele Krankheiten im höheren Alter häufiger auftreten, wird dieser Anstieg zu einer Zunahme der Zahl an Menschen mit chronischen Krankheiten führen, insbesondere in der Gruppe der über 80-Jährigen. Dementsprechend ist auch mit einem Anstieg der Behandlungen und Kosten zu rechnen.

In Bezug auf die Gesamtbevölkerungszahlen sehen die Prognosen für Deutschland anders aus als für die Schweiz und Österreich (s. Tab. 2.5). Hier soll die Bevölkerung bis zum Jahr 2030 um 5 % abnehmen. In der Schweiz geht man dagegen von einer weiteren Zunahme um 11 % aus, in Österreich von einer Zunahme um 12 %. Angesichts der hohen Anzahl an Flüchtlingen aus dem Nahen Osten werden diese Prognosen allerdings inzwischen als unsicher eingestuft.

Tab. 2.5: Prognosen zur Bevölkerungsentwicklung in Deutschland, Österreich und in der Schweiz.

Jahr	2010	2020	2030
Deutschland			
Gesamtbevölkerung (in Mio.)	81,5	79,9 (–2,0 %)	77,4 (–5,0 %)
Personen älter als 65 Jahre (in %)	20,6	23,3 (+13,1 %)	28,8 (+39,8 %)
Österreich			
Gesamtbevölkerung (in Mio.)	8,39	9,12 (+8,7 %)	9,43 (+12,4 %)
Personen älter als 65 Jahre (in %)	17,7	19,0 (+7,3 %)	22,8 (+28,8 %)
Schweiz			
Gesamtbevölkerung (in Mio.)	7,86	8,40 (+6,9 %)	8,74 (+11,2 %)
Personen älter als 65 Jahre (in %)	17,1	20,1 (+17,5 %)	24,2 (+41,5 %)

Quellen: Statistisches Bundesamt. Bevölkerung Deutschlands bis 2060, Ergebnisse der 12. Koordinierten Bevölkerungsvorausberechnung (www.destatis.de); Bundesamt für Statistik. Szenarien zur Bevölkerungsentwicklung der Schweiz: 2010–2060. Neuchâtel 2010 (www.bfs.admin.ch); Statistik Austria. Vorausberechnete Bevölkerungsstruktur für Österreich 2015–2100 laut Hauptszenario (www.statistik.at).

2.2.3 Sterbefälle und Mortalitätsraten

Die Mortalitäts- oder Sterberate für ein bestimmtes Gebiet in einem definierten Kalenderjahr wird aus dem Verhältnis zwischen der Anzahl an Sterbefällen und der ständigen Einwohnerzahl in der Mitte des Jahres berechnet und meist pro 100.000 Einwohner angegeben.

$$\text{Mortalitätsrate im Jahr X (pro 100.000)} = \frac{\text{Anzahl der Sterbefälle im Jahr X}}{\text{Bevölkerungszahl in der Jahresmitte von X}} \cdot 100.000$$

So starben z. B. in Österreich im Jahr 2014 78.252 Personen. Bei einer Bevölkerungszahl von 8,58 Mio. Einwohnern ergibt sich daraus eine Mortalitätsrate von 911,5 Sterbefällen pro 100.000 Einwohner. Analog werden die geschlechts- und altersspezifischen Mortalitätsraten berechnet. In diesem Fall werden im Zähler und im Nenner nur die Zahlen des jeweiligen Geschlechts bzw. der jeweiligen Altersgruppe eingesetzt. Für die Berechnung der Mortalitätsrate der 50- bis 54-jährigen Männer wird somit im Zähler die Zahl der Todesfälle der Männer im Alter von 50–54 Jahren und im Nenner die Zahl der Männer dieses Alters in der ständigen Wohnbevölkerung des betrachteten Gebietes zur Jahresmitte verwendet.

Abb. 2.11 zeigt die auf diese Weise berechneten altersspezifischen Mortalitätsraten auf einer logarithmischen Skala für Männer und Frauen in der Schweiz in den Jahren 1900 und 2014. Ganz offensichtlich sind die Mortalitätsraten 2014 deutlich niedriger als im Jahr 1900. Bei den 40-Jährigen ist die Sterberate heute zehnmal niedriger, und bei Kindern im Alter von 2–15 Jahren ist die Reduktion noch deutlicher. Die Abbildung zeigt auch, dass im Jahr 2014 die Mortalitätsraten bei den über 20-jährigen Männern höher liegen als bei den gleichaltrigen Frauen. Besonders ausgeprägt ist dies im Alter zwischen 20 und 35 Jahren. Ein ähnliches Bild zeigt sich in Österreich und in West- bzw. Ostdeutschland in den Jahren 1960 und 2013 (s. Abbildungen in Kap. 2.2 auf unserer Lehrbuch-Homepage).

Die Sterblichkeit von Neugeborenen wird anders berechnet: Die Zahl aller in einem Kalenderjahr innerhalb des ersten Lebensjahres verstorbenen Kinder wird

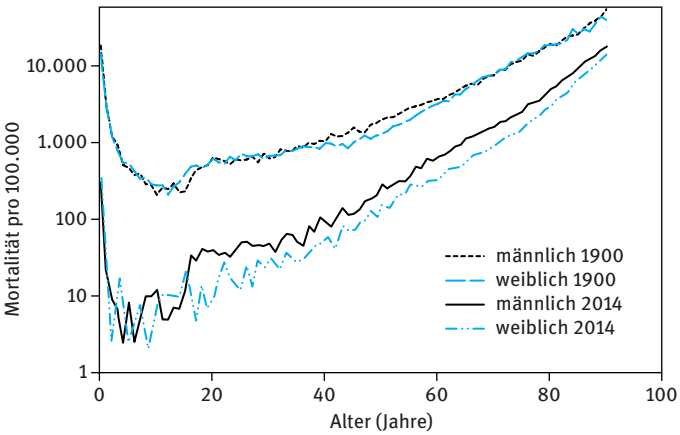

Abb. 2.11: Mortalitätsraten pro 100.000 Personen in der Schweiz in den Jahren 1900 und 2014, berechnet nach Altersgruppen und Geschlecht. Die y-Achse weist eine logarithmische Skala auf (Quelle: The Human Mortality Database; www.mortality.org).

zur Anzahl der in diesem Jahr lebend geborenen Kinder ins Verhältnis gesetzt. Man bezeichnet diese Größe als *Säuglingssterblichkeit*. In der Regel wird sie pro 1.000 Lebendgeborene angegeben.

$$\text{Säuglingssterblichkeit (pro 1.000)} = \frac{\text{Anzahl der Sterbefälle im ersten Lebensjahr}}{\text{Anzahl lebend geborener Kinder}} \cdot 1.000$$

Die Säuglingssterblichkeit ist bei den Jungen etwas höher als bei den Mädchen. Ebenso wie in allen anderen, sich wirtschaftlich erfolgreich entwickelnden Ländern ist sie in Deutschland, Österreich und der Schweiz in den vergangenen Jahren kontinuierlich gesunken (s. Abbildung in Kap. 2.2 auf unserer Lehrbuch-Homepage). Im Jahr 2014 betrug sie etwa 4,5 pro 1.000 lebend geborener Jungen und 3,5 pro 1.000 lebend geborener Mädchen. Die *Kindersterblichkeit* beziffert die Anzahl der Kinder, die im Zeitraum der ersten fünf Lebensjahre sterben, und die *neonatale Sterblichkeit* die Anzahl der Kinder, die innerhalb von 28 Tagen nach Geburt versterben. Beide Größen werden wiederum auf 1.000 Lebendgeburten bezogen.

2.2.4 Lebenserwartung

In einem Gedankenexperiment kann man sich 1.000 lebend geborene Mädchen vorstellen und sich fragen, wie viele von ihnen den ersten Geburtstag feiern könnten, wenn für sie die Säuglingssterblichkeit für Mädchen aus dem Jahr 1900 in der Schweiz gelten würde. Weiterhin könnte man sich fragen, wie viele von den Mädchen, die ein Jahr alt geworden wären, ihren zweiten Geburtstag feiern würden, wenn die Mortalitätsrate der 1- bis 2-jährigen Mädchen aus dem Jahr 1900 gelten würde. Solche Berechnungsschritte kann man für jedes weitere Lebensjahr machen. Aus den Prozentsätzen dieser hypothetischen Personen, die den jeweiligen Geburtstag erleben, lässt sich eine Überlebenskurve zeichnen (Abb. 2.12). Etwa 20 % der lebend geborenen Mädchen würden vor dem 5. Lebensjahr sterben und nur rund 40 % dieser Mädchen würden den 60. Geburtstag feiern können.

Jeder Mensch wird einmal sterben. Wir können nun das jeweilige Sterbealter erfassen und den Mittelwert des Sterbealters aller Personen berechnen. Das Ergebnis nennt man die *durchschnittliche Lebenserwartung ab Geburt*. In unserer Berechnung mit den Mortalitätsraten von 1900 resultiert daraus für Mädchen eine Lebenserwartung von 47,8 Jahren. Sie entspricht der durchschnittlichen Zahl an zu erwartenden Lebensjahren unter der Voraussetzung, dass die in einem bestimmten Jahr beobachteten altersspezifischen Mortalitätsraten für das ganze Leben gelten würden. Die Lebenserwartung lässt sich auch einfach grafisch finden: Es genügt ein Rechteck, das die gleiche Fläche aufweist wie die Fläche unter der Überlebenskurve (Abb. 2.12). Das Rechteck zeigt eine hypothetische Überlebenskurve, bei der alle Personen bis zu einem gewissen Alter überleben und dann im selben Alter sterben.

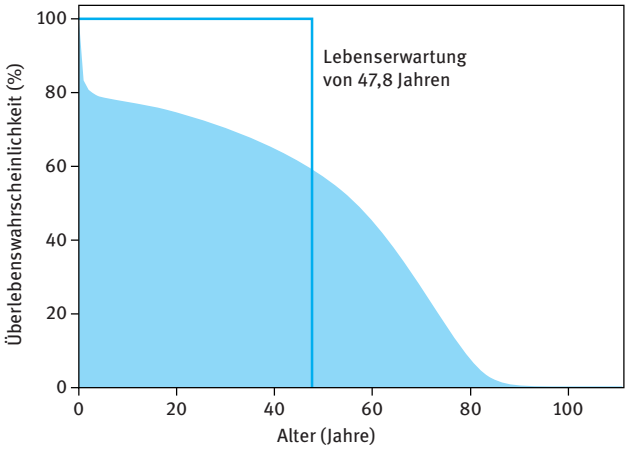

Abb. 2.12: Überlebenskurve einer hypothetischen Gruppe von lebend geborenen Mädchen, die mit den Schweizer Mortalitätsraten von 1900 versterben würden. Die grün gefärbte Fläche unter der Überlebenskurve entspricht der Fläche des eingezeichneten Rechtecks.

Die Fläche unter der Überlebenskurve und damit die Lebenserwartung erhöhen sich deutlich, sobald die Kindersterblichkeit sinkt. Genau das geschah während des letzten Jahrhunderts in den Industrieländern. Die altersspezifischen Mortalitätsraten sanken in allen Altersgruppen kontinuierlich ab. Dies führte zu einer Erhöhung der Lebenserwartung ab Geburt in der Schweiz (s. Abbildung in Kap. 2.2 auf unserer Lehrbuch-Homepage), in Österreich und in Deutschland (Abb. 2.13). So betrug die Lebenserwartung ab Geburt in Österreich um 1900 noch 43,4 Jahre für Frauen und 40,6 Jahre für Männer. Bis 1950 stieg sie auf 67,3 J. (♀) bzw. 62,2 J. (♂) an, und bis 2014 dann noch einmal auf 83,7 J.(♀) bzw. 78,9 J. (♂). Ähnlich sieht es in der Schweiz aus, wo die Lebenserwartung von 48,8 Jahren für Frauen und 46,1 Jahren für Männer im Jahr 1900 auf 74,1 J. (♀) bzw. 68,7 J. (♂) im Jahr 1960 und weiter auf 85,1 J. (♀) bzw. 80,9 J. (♂) im Jahr 2014 anstieg. Im Jahr 1918 führte eine Grippepandemie nicht nur in Europa zu einem Einbruch in der Lebenserwartung. Hieran wird deutlich, dass es sich bei der Lebenserwartung um eine hypothetische Konstruktion handelt. Sie zeigt die Überlebenskurven von Personen, die in jedem Altersjahr die Mortalitätsrate durchleben müssten, wie sie zur Zeit ihrer Geburt herrschte. Die hypothetische Gruppe von Personen, die während der Grippepandemie geborenen wurden, würde somit ihr ganzes Leben in der Situation der Grippepandemie von 1918 leben.

Seit dem zweiten Weltkrieg haben Frauen in West- und Ostdeutschland gegenüber Männern eine etwa um 5 Jahre höhere Lebenserwartung. Wie in der Schweiz und in Österreich zeigt die zeitliche Entwicklung für Frauen und Männer in Westdeutschland eine stetige Zunahme der Lebenserwartung. In Ostdeutschland war hingegen von 1980 bis zur Wiedervereinigung 1990 eine Verlangsamung dieses Trends

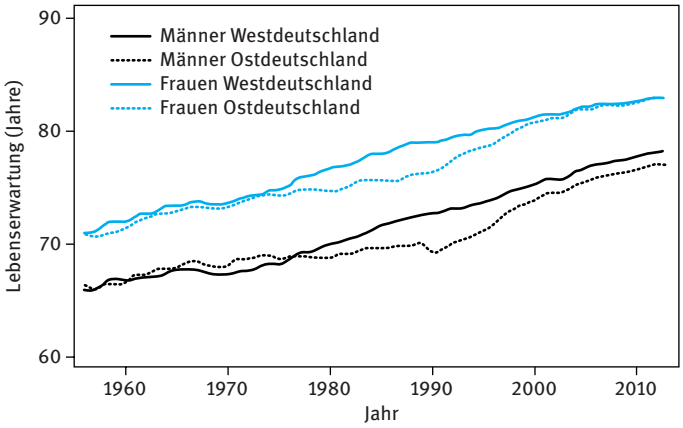

Abb. 2.13: Durchschnittliche Lebenserwartung ab Geburt in Deutschland zwischen 1956 und 2013 (Quelle: The Human Mortality Database; www.mortality.org).

festzustellen. In der Zwischenzeit haben sich die Lebenserwartungen in West- und Ostdeutschland einander angenähert.

Die mittlere Lebenserwartung bei Geburt ist eine der zentralen internationalen Vergleichsziffern im Gesundheitswesen. Sie spiegelt die sozioökonomischen und gesundheitlichen Lebensverhältnisse in einer Gesellschaft zu einem bestimmten Zeitpunkt wider. Vergleicht man die mittlere Lebenserwartung bei Geburt in verschiedenen Gesellschaften bzw. innerhalb einer Gesellschaft zu verschiedenen Zeitpunkten, lassen sich daraus wesentliche Rückschlüsse auf das allgemeine Entwicklungsniveau einer Bevölkerung ziehen.

2.2.5 Todesursachen und potentiell verlorene Lebensjahre

Wie bereits dargestellt, steigen die Sterberaten für Männer und Frauen mit zunehmendem Lebensalter unterschiedlich stark an. Aus den Sterberaten und der urnenförmigen Altersstruktur, wie sie heute für die Schweiz, Österreich und Deutschland vorliegt, ergibt sich die Altersverteilung der Todesfälle in den drei Ländern. Über die letzten hundert Jahre hat sich diese Verteilung stark verändert. 2014 traten in der Schweiz weniger als 1 % aller Todesfälle bei Kindern und Jugendlichen auf, die jünger als 15 Jahre alt waren. Im Jahr 1900 waren hingegen noch 32,2 % der Gestorbenen jünger als 15 Jahre! Dagegen traten im Jahr 2014 über 40 % der Todesfälle bei Menschen auf, die 85 Jahre oder älter waren, im Jahr 1900 waren es nur 2,1 %. Auch die Verteilung nach Todesursachen hat sich stark verändert (s. eine Box in Kap. 2.2 auf unserer Lehrbuch-Homepage).

Um deutlich zu machen, dass bestimmte Todesursachen bei jüngeren Personen eine wichtige Rolle spielen, hat man die *Zahl an potentiell verlorenen Lebensjahren*, aufgeschlüsselt nach Todesursachen, als weitere Kennziffer berechnet. In der englischen Terminologie werden sie *Potential Years of Life Lost* (PYLL) genannt. Für jede Todesursache lässt sich pro Kalenderjahr berechnen, wie viele Menschen hieran vor dem 70. Lebensjahr verstorben sind. Wenn jemand beispielsweise im Alter von 30 Jahren an einem Verkehrsunfall stirbt, dann gehen durch diesen Todesfall 40 potentielle Lebensjahre verloren. Man zählt nun alle potentiell verlorenen Lebensjahre der an einer Todesursache verstorbenen Personen zusammen. Diese Zahlen sowie ihre prozentuale Verteilung nach Todesursachen dienen als Hinweis auf das Präventionspotential für einzelne, zum Tod führende Krankheiten. Betrachtet man z. B. die relative Verteilung nach Krankheitsgruppen in Deutschland, so gehen bei den Frauen 41 % der potentiell verlorenen Lebensjahre auf Krebserkrankungen zurück (Männer: 25 %), 7,8 % auf Unfälle (Männer: 13,1 %) und 5,1 % auf Suizide (Männer: 9,3 %).

Ein weiterer Begriff, der in diesem Zusammenhang häufig genannt wird, ist die „gesunde Lebenserwartung“ (*Healthy Life Expectancy*). Diese wird definiert als die durchschnittliche Anzahl an zu erwartenden Lebensjahren, die bei guter Gesundheit bzw. ohne nachhaltige Behinderung verbracht werden. Schließlich wird die Krankheitslast (*Burden of Disease*) in *Disability Adjusted Life Years* (DALYs) angegeben, wobei ein DALY einem durch Erkrankung oder vorzeitigen Tod verlorenen gesunden Lebensjahr entspricht (s. Kap. 10.1.2).

2.3 Biostatistik

Marcel Zwahlen

Wir lesen in einem Fachartikel, dass bei einer bestimmten Therapieform von 100 Behandelten nur halb so viele versterben wie bei einer anderen Form der Therapie. Ist dieser Unterschied statistisch gut abgesichert (*statistisch signifikant*)? Oder ist es möglich, dass er nur auf Zufall beruht? Es könnte z. B. sein, dass in der ersten Gruppe eine Person verstarb, in der zweiten jedoch zwei. In der ersten Gruppe starben damit tatsächlich nur halb so viele Menschen wie in der zweiten Gruppe. Wie stark unterscheidet sich der Therapieerfolg bei diesen beiden Behandlungsformen nun wirklich? Mit Hilfe der Statistik versuchen wir, über numerische Informationen Antworten auf solche Fragen zu erhalten. Statistik befasst sich mit dem Sammeln, Zusammenfassen, Darstellen und Interpretieren von Daten. *Biostatistik* ist der Zweig der Statistik, der diese Aufgaben in der Biomedizin und in Public Health übernommen hat.

2.3.1 Warum brauchen wir Statistik?

„In God we trust. All others must have data."
W.E. Demming (1900–1993; amerik. Physiker und Statistiker)

„It is easy to lie with statistics. It is hard to tell the truth without statistics."
A. Dunkels (1939–1998; schwed. Mathematiker und Lehrer)

Statistik und statistische Verfahren dienen dazu, aus Situationen, die typischerweise mit einer gewissen Variabilität auftreten, möglichst wahrheitsgemäße Schlüsse zu ziehen. Insbesondere biologische Prozesse zeigen oft eine solche inhärente Variabilität. Dies spiegelt sich dann auch in biomedizinischen Messwerten wider. So variiert beispielsweise der arterielle Blutdruck nicht nur von Mensch zu Mensch, sondern auch bei einem Individuum von Stunde zu Stunde. In einer Population von Individuen äußert sich Variabilität in Form von zufällig auftretenden Ereignissen oder Messwerten. Einerseits können beispielsweise Personen, die gegen eine bestimmte Infektionskrankheit geimpft wurden, trotz Impfung an dieser Infektion erkranken, andererseits können ungeimpfte Personen gesund bleiben. Wenn wir diese Situation aus statistischer Sicht betrachten, stellen sich uns u. a. folgende Fragen:

- Was kann daraus geschlossen werden, wenn bei den geimpften Personen ein größerer Anteil gesund bleibt als bei den ungeimpften?
- Wie wirksam ist der Impfstoff? Ist der Unterschied zwischen Geimpften und Ungeimpften vielleicht zufällig zustande gekommen?
- Gaukelt uns eine Verzerrung bei der Studienpopulation möglicherweise eine Wirkung der Impfung nur vor? So konnte z. B. die Gruppe der Geimpften mehr Interesse an präventiven Maßnahmen gezeigt haben als die der Ungeimpften. Damit wäre denkbar, dass sich beide Gruppen im Gesundheitsverhalten und in den generellen Lebensumstanden unterscheiden. Dies alles sind Faktoren, die die Erkrankungswahrscheinlichkeit beeinflussen konnten.

Statistische Methoden erlauben es, die ersten beiden Fragen zu beantworten. Das in der dritten Frage angesprochene Problem eines Selektionsbias (s. Kap. 2.1.8) kann durch eine sorgfaltige Planung der durchzuführenden Studie verhindert werden.

Die *Hauptarbeitsbereiche der Biostatistik* sind

- die Mithilfe bei der Planung von Studien (s. die verschiedenen Studientypen in Kap. 2.1)
- die Beschreibung und Zusammenfassung von erhobenen Daten (z. B. des mittleren Blutdrucks in einer Population, s. „deskriptive Statistik")

- die Quantifizierung von wichtigen Kenngrößen in Populationen oder Patientengruppen (z. B. die Inzidenz einer Infektion, s. „Schätzen von Parametern“)
- das Testen von präzisen quantitativen Hypothesen („Impfstoff A ist 20 % wirksamer als Impfstoff B“)

2.3.2 Klassifikation von Daten

Um in der Biomedizin und in Public Health Antworten auf Fragen zu bekommen, werden in der Regel Studien durchgeführt, die Messungen beinhalten. Gemessen werden bestimmte Charakteristika (**Variablen**), die Antworten auf die bestehenden Fragen versprechen. Häufig sind dies Untersuchungen bei StudienteilnehmerInnen. Es kann sich aber auch um Messwerte handeln, die an Versuchstieren gewonnen wurden oder um Charakteristika von Krankenhäusern oder Analyseergebnisse aus Urinproben. Jeder Aspekt, der untersucht wird, wie etwa der Blutdruck, der Cholesterinspiegel oder das Geschlecht, entspricht in der Regel einer Variablen. Bevor die Anwendung bestimmter statistischer Verfahren festgelegt und erste Berechnungen durchgeführt werden, lohnt es sich, die vorhandenen Daten anzusehen und sie nach Datentypen zu ordnen. In einem ersten Schritt wird zwischen quantitativer und kategorischer Information unterschieden.

Quantitative Daten sind entweder *kontinuierliche* oder *diskrete Daten*. Als kontinuierliche Variable bezeichnet man einen Messwert, der sich auf einer kontinuierlichen Skala mit einer definierten Maßeinheit abbilden lässt. Kontinuierliche Variablen sind z. B. das Körpergewicht oder ein Cholesterinwert. Sie können jeden beliebigen Wert auf der Skala des Messgerätes einnehmen. Im Gegensatz dazu kann eine diskrete Variable nur eine beschränkte Anzahl, meist ganzzahliger Werte annehmen. Beispiele hierfür sind die Anzahl von Geburten oder von Krankenhausaufenthalten im letzten Jahr.

Kategorische Daten werden auch *nominale* oder *qualitative Daten* genannt. Hierbei handelt es sich um nicht-numerische Daten, wie beispielsweise der Geburtsort, die Nationalität, die Augenfarbe oder die Art eines Medikaments. Eine wichtige Untergruppe kategorischer Daten sind so genannt *binäre oder dichotome Variablen,* die nur zwei mögliche Werte kennen. So ist das Geschlecht entweder weiblich oder männlich, und der Teilnehmer an einer Studie ist bei Studienende entweder am Leben oder gestorben.

Bei **geordneten kategorischen Daten** gehen wir davon aus, dass den Kategorien – auch wenn sie nicht-numerischer Art sind – eine natürliche Ordnung zukommt. Geordnete kategorische Daten sind z. B. die Antworten auf die folgende Frage:

„Während meines Krankenhausaufenthaltes wurde ich mit Respekt und Würde behandelt.“

Bitte beantworten Sie, ob Sie dieser Aussage
a) überhaupt nicht zustimmen
b) ein wenig zustimmen
c) stark zustimmen
d) vollumfänglich zustimmen

Ein weiteres Beispiel für eine solche natürliche Ordnung sind die Stadien einer Krebserkrankung: Stadium I hat eine bessere Prognose als Stadium IV.

2.3.3 Transparentes Zusammenfassen der erhobenen Daten

Die quantitativen Daten, die in einer Studie erhoben wurden, müssen in einem ersten Schritt geeignet zusammengefasst werden, um eine bessere Übersichtlichkeit zu erreichen.

Betrachten Sie z. B. die folgende Situation:

In einer Studie, an der 200 Personen teilnahmen, wurden u. a. Gewicht und Körpergröße gemessen. Anhand dieser Werte wurde anschließend der Body-Mass-Index (BMI) der TeilnehmerInnen durch Division von Körpermasse (in Kilogramm) durch das Quadrat der Körpergröße (in Metern) berechnet. Die alleinige Auflistung der 200 BMI-Werte wäre nun bei der Beurteilung dieser Daten wenig hilfreich:

BMI-Werte [kg/m^2] der Personen 1–10:
24,0; 27,6; 28,7; 29,0; 25,4; 25,7; 27,8; 25,3; 28,4; 29,0

BMI-Werte [kg/m^2] der Personen 191–200:
28,5; 24,8; 28,6; 21,8; 24,4; 24,4; 21,3; 26,8; 27,7; 22,9

Es ist sinnvoller, eine leicht verständliche Zusammenfassung dieser Werte zu erstellen. Das kann mittels *grafischer Darstellung* oder mit Hilfe geeigneter *Kennzahlen* geschehen. Nützlich ist in diesem Zusammenhang die so genannte **Fünf-Zahlen-Zusammenfassung** (*Five-Number Summary*). Zu diesen fünf Zahlen gehören:
- **Tiefster Wert** (Minimum)
- **Unteres Quartil**: Der Wert, der die vorliegende Reihe von Werten so unterteilt, dass 25 % der Werte kleiner als dieser Wert sind.
- **Median** (m): Der Wert, der die Reihe so unterteilt, dass (höchstens) die Hälfte der Werte kleiner als m und (höchstens) die Hälfte der Werte größer als m sind. Bei einer geraden Anzahl von Werten (k = Anzahl der vorliegenden Werte) wird die Mitte zwischen dem (k/2)-ten und (k/2 + 1)-ten Wert genommen.
- **Oberes Quartil**: Der Wert, der die Reihe von Werten so unterteilt, dass 75 % der Werte kleiner als das obere Quartil sind.
- **Höchster Wert** (Maximum).

Bei den 200 Personen, für die der BMI berechnet wurde, ergäben sich daraus z. B. die folgenden Werte (in kg/m²) der Fünf-Zahlen-Zusammenfassung:

> Minimum: 16,70; Unteres Quartil: 23,50; Median: 25,10; Oberes Quartil: 26,85; Maximum: 39,20.

Diese fünf Kennzahlen lassen sich auch in einem so genannten **Boxplot** (Kastengrafik) oder *Box-Whisker-Plot* darstellen (Abb. 2.14). Im Boxplot sehen wir in der Mitte eine dunkler eingefärbte Box, welche durch die Werte des unteren und oberen Quartils begrenzt ist. 50 % aller Werte liegen innerhalb des Interquartilbereichs zwischen 23,50 und 26,85. In der Mitte dieser Box ist der Median-Wert eingezeichnet. Die beiden Linien, die von den Rändern der Box ausgehen, werden *Whisker* (Antennen, Fühler) genannt. Die Länge dieser Whisker ist auf das 1,5-Fache des Interquartilabstands beschränkt. In unserem Beispiel beträgt der Interquartilabstand 26,85 − 23,5 = 3,35. Die Begrenzung des oberen Whiskers liegt also maximal bei 26,85 + (1,5 × 3,35) = 31,875. Der untere Whisker erstreckt sich bis maximal 23,5 − (1,5 × 3,35) = 18,475. Werte, die weiter als die beiden Whisker vom Median entfernt liegen, werden einzeln dargestellt und als Ausreißerwerte bezeichnet.

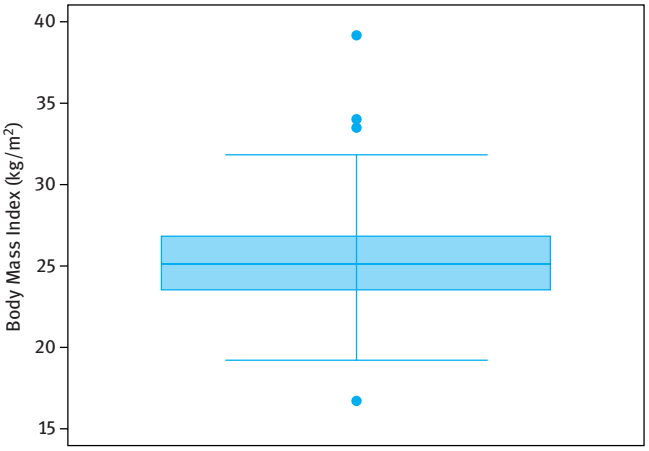

Abb. 2.14: Boxplot-Darstellung der BMI-Werte der 200 Personen aus dem Anwendungsbeispiel. Bei den Punkten außerhalb der Whisker handelt es sich um so genannte Ausreißerwerte.

Eine andere Form der grafischen Darstellung ist das **Histogramm**. Hierbei werden zuerst Werteintervalle gebildet, anschließend wird gezählt, wie viele der vorliegenden Werte in die jeweiligen Intervalle fallen. Das Ergebnis kann man dann auf zwei verschiedene Arten grafisch darstellen. Entweder wird die Anzahl oder der Prozentsatz der Werte aufgezeigt, die jeweils in die gebildeten Werteintervalle fallen. Abb. 2.15 zeigt eine solche Darstellung. Die gewählten Intervalle haben hier eine Länge von

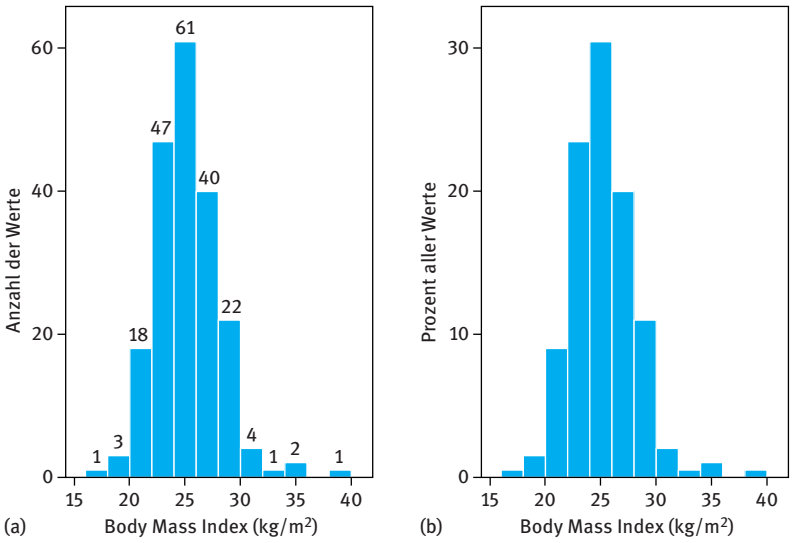

Abb. 2.15: Histogramm der BMI-Werte der 200 Personen aus dem Anwendungsbeispiel.
(a) Histogramm, bei dem die Anzahl der Werte in der jeweiligen Wertegruppe angegeben sind.
(b) Histogramm, das die jeweiligen Prozentsätze angibt.

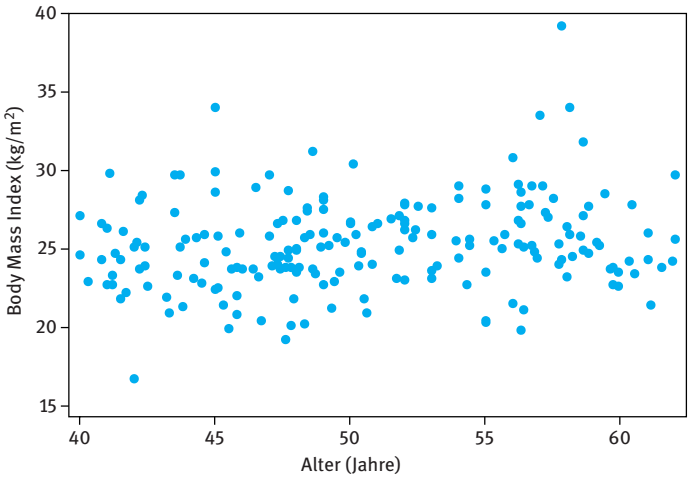

Abb. 2.16: Streudiagramm (Scatter Plot), das das Alter der 200 Personen aus dem Anwendungsbeispiel zu ihrem Body Mass Index in Relation setzt.

2 kg/m² (z. B. 16 bis < 18 kg/m², 18 bis < 20 kg/m² etc.). Der Nachteil eines solchen Histogramms ist, dass es von der gewählten Intervalleinteilung abhängt, welches Bild man erhält.

Eine weitere Möglichkeit der grafischen Darstellung ist das **Streudiagramm**. Hierdurch können zwei verschiedene Merkmale gleichzeitig dargestellt werden. In unserem Anwendungsbeispiel ließe sich auf diese Weise etwa der BMI mit dem Alter der 200 untersuchten Personen verknüpfen (Abb. 2.16).

Oft werden auch der **Mittelwert** und die **Standardabweichung** als Kennzahlen zur Zusammenfassung der vorliegenden Werte verwendet. Die Anzahl der Werte des Datensatzes bezeichnet man hierbei mit N.

Formel 2.1: Formeln für die Berechnung des Mittelwertes und der Standardabweichung einer Wertereihe. SD = Standard Deviation (engl.).

$$\text{Mittelwert} = \frac{X_1 + \ldots + X_N}{N} = \frac{\sum_{j=1}^{j=N} X_j}{N}$$

$$\text{Standardabweichung (SD)} = \sqrt{\frac{\sum_{j=1}^{j=N} (X_j - \text{Mittelwert})^2}{N - 1}}$$

Der Mittelwert ist eine Kennzahl für typische Werte in der Mitte einer Datenreihe, die Standardabweichung kennzeichnet dagegen die Variabilität der betrachteten Werte. Zu beachten ist, dass bei der Berechnung der Standardabweichung die Summe der quadrierten Abstände zum Mittelwert durch die um 1 reduzierte Anzahl der Werte geteilt wird (N–1).

Ständen uns alle 200 BMI-Werte aus unserem Anwendungsbeispiel zur Verfügung, ließe sich daraus ein Mittelwert von 25,3 kg/m² sowie eine Standardabweichung von 2,92 kg/m² berechnen. In unserem Beispiel nimmt die Standardabweichung damit einen ähnlichen Wert wie der Interquartilabstand ein, der 3,35 kg/m² betrug.

Mittelwert und Standardabweichung reagieren empfindlich darauf, wenn einige wenige Werte weit außerhalb des übrigen Wertebereichs liegen. So würde sich die Standardabweichung z. B. von 2,92 auf 4,98 kg/m² erhöhen, wenn unter den Werten unseres Beispiels anstatt der zehn höchsten BMI-Werte zwischen 29,8 kg/m² und 39,2 kg/m² zehn Werte von jeweils 45 kg/m² gewesen wären. Der Mittelwert würde nun 25,9 kg/m² betragen. Median und Interquartilbereich würden sich jedoch nicht ändern. Tab. 2.6 fasst die *Vor- und Nachteile der Kennzahlen quantitativer Daten* zusammen.

Tab. 2.6: Vor- und Nachteile der Kennzahlen quantitativer Daten.

	Vorteil	Nachteil
Kennzahlen für die Mitte		
Mittelwert	Einfach zu berechnen, gute statistische Eigenschaften	Reagiert empfindlich auf Ausreißerwerte
Median	Einfach zu verstehen, reagiert nicht sensibel auf Ausreißerwerte (= robust gegenüber Ausreißerwerten)	Hat komplexe statistische Eigenschaften
Kennzahlen für die Variabilität		
Standardabweichung	Hat gut verstandene statistische Eigenschaften	Ist kompliziert zu berechnen, reagiert empfindlich auf Ausreißerwerte
Interquartilbereich	Einfach zu verstehen: 50 % aller Werte liegen in diesem zentralen Bereich	Hat komplexe statistische Eigenschaften

Will man dagegen die Resultate von **qualitativen Daten** zusammenfassen, ist es nicht sinnvoll, Median oder Mittelwert zu berechnen. Dies gilt auch dann, wenn Zahlencodes verwendet wurden, wie z. B. die Zahlen 1 bis 5 zur Kodierung des Personenstands (schweizerisch: Zivilstand) in ledig, verheiratet, geschieden, verwitwet, getrennt lebend. Eine nützliche Information bei qualitativen Daten ist die *prozentuale Verteilung* auf die verschiedenen Kategorien. Diese kann dann anhand einer **Tabelle** (Tab 2.7) oder grafisch in Form eines **Kuchendiagramms** (*Pie chart*; Abb. 2.17) oder eines **Häufigkeitsdiagramms** (s. Abb. 2.18) dargestellt werden. Die Kuchengrafik bezeichnet man auch als *Kreisdiagramm*, die Häufigkeitsgrafik als *Balkendiagramm* (*Bar chart*).

Tab. 2.7: Zivilstand (Personenstand) der 30- bis 49-jährigen Männer und Frauen in der Schweiz (Schweizerische Gesundheitsbefragung 2007).

Zivilstand	Männer	Frauen
Ledig	26,5 %	18,5 %
Verheiratet	64,8 %	68,7 %
Verwitwet	0,6 %	1,4 %
Geschieden	6,3 %	9,4 %
Getrennt lebend	1,8 %	2,0 %
Gesamt	100 %	100 %

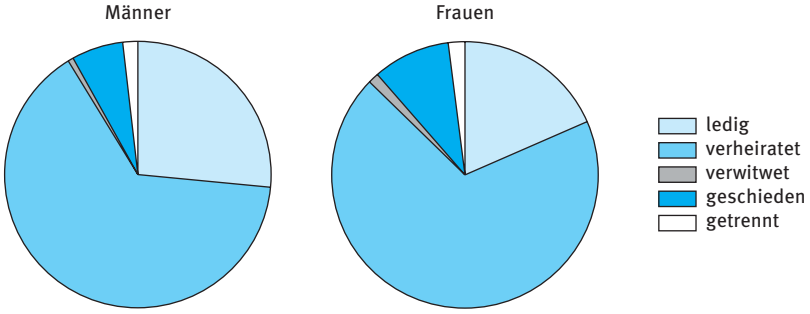

Abb. 2.17: Kuchengrafik, die den Zivilstand (Personenstand) der 30- bis 49-jährigen Männer und Frauen in der Schweiz wiedergibt (Schweizerische Gesundheitsbefragung 2007; die genauen Prozentsätze zeigt Tab. 2.7).

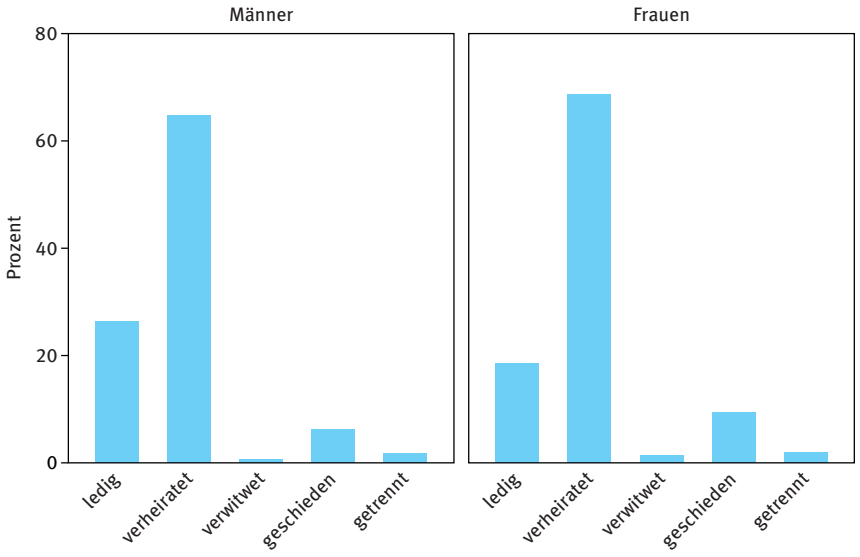

Abb. 2.18: Häufigkeitsgrafik, die den Zivilstand (Personenstand) der 30- bis 49-jährigen Männer und Frauen in der Schweiz wiedergibt (Daten aus der Schweizerischen Gesundheitsbefragung 2007).

2.3.4 Variabilität des Mittelwertes bei wiederholten Zufalls-Stichproben

Da es nur ausnahmsweise möglich ist, Untersuchungen ganzer **Populationen** durchzuführen, ist man in der Regel dazu gezwungen, sich mit der Analyse einer Teilmenge einer Population zu begnügen. Wenn diese Teilmenge nach dem Zufallsprinzip ausgewählt wurde, wird sie als *zufällig gezogene Stichprobe* bezeichnet. Die Wahrscheinlichkeitslehre erlaubt es nun, anhand einer solchen Stichprobe Aussagen darüber zu machen, wie sich die statistischen Kennzahlen der zufällig gezogenen Stichprobe

von den wahren Werten in der Gesamtpopulation unterscheiden. Wenn man anhand der Resultate einer Stichprobe Aussagen über die ganze Population machen will, muss man allerdings berücksichtigen, dass aufgrund des Zufalls mehrere, nach dem gleichen Zufallsprinzip gezogene Stichproben unterschiedliche Kennzahlen liefern. Dieses Phänomen der „Stichprobenvariation" soll nun anhand von Computersimulationen illustriert werden.

Computersimulation der Stichprobenvariabilität

Hierzu stellen wir uns vor, dass wir im Jahr 2007 in der Schweiz alle rund 2,4 Mio. Personen im Alter zwischen 30 und 49 Jahren nach ihrem Personenstand (Zivilstand) befragt hätten. Es zeigte sich, dass exakt 60 % der Befragten verheiratet waren. Wir ziehen nun am Computer eine zufällige Stichprobe von einer bestimmten Größe (z. B. 50 Personen) aus der Gesamtpopulation und berechnen anschließend den Prozentsatz der verheirateten Personen aus dieser Stichprobe. Das Ganze wird 10.000-mal wiederholt, und zum Schluss wird die Verteilung der in den Stichproben berechneten Prozentsätze mittels eines Histogramms beschrieben. Dieses Prozedere wird dann für eine Stichprobengröße von 100, 300 und 500 Personen wiederholt (Abb. 2.19). Es wird bei allen Stichprobengrößen deutlich, dass die Verteilung des berechneten Prozentsatzes jeweils um die Mitte, den wahren Wert von 60 % schwankt. Allerdings variieren die Resultate bei einer Stichprobengröße von 50 Personen relativ stark zwischen 40 % und 80 %. Dagegen kommt es bei einer Stichprobengröße von 300 Personen kaum vor, dass der berechnete Prozentsatz kleiner als 50 % oder größer als 70 % wird. Je mehr Personen eine Stichprobe umfasst, desto weniger variieren also die in der Stichprobe berechneten Resultate. Im Grenzfall einer Vollerhebung entspricht der berechnete Wert exakt dem wahren Wert.

Auch für die Berechnung des Mittelwertes einer Stichprobe gilt, dass der hier berechnete Wert vom wahren Mittelwert umso weniger abweicht, je größer die Stichprobe ist. Bei unserem Beispiel der rund 2,4 Mio. SchweizerInnen im Alter zwischen 30 und 49 Jahren sind die Body-Maß-Index-Werte normalverteilt mit einem Mittelwert von 25 kg/m^2 und einer Standardabweichung von 4 kg/m^2. Werden nun aus der Gesamtpopulation wiederholt Stichproben verschiedener Größe gezogen und wird pro Stichprobe der Mittelwert des BMI berechnet, so ergibt sich hieraus eine Verteilung der für die Stichproben berechneten Mittelwerte um die Mitte, den wahren Wert von 25 kg/m^2 herum. Die Mittelwerte variieren dabei umso mehr, je kleiner die Anzahl an Personen in der Stichprobe ist (s. Abbildung in Kap. 2.3 auf unserer Lehrbuch-Homepage).

Aus der Wahrscheinlichkeitslehre ergibt sich auch, dass die Stichprobenvariabilität einer berechneten Proportion oder eines berechneten Mittelwertes annährungsweise durch die so genannte *Normalverteilung* beschrieben werden kann, wenn die Stichproben nach dem Zufallsprinzip gezogen wurden. Je größer hierbei die Stichprobengröße N ist, desto exakter stimmt diese Annäherung. Die normalverteilten Werte

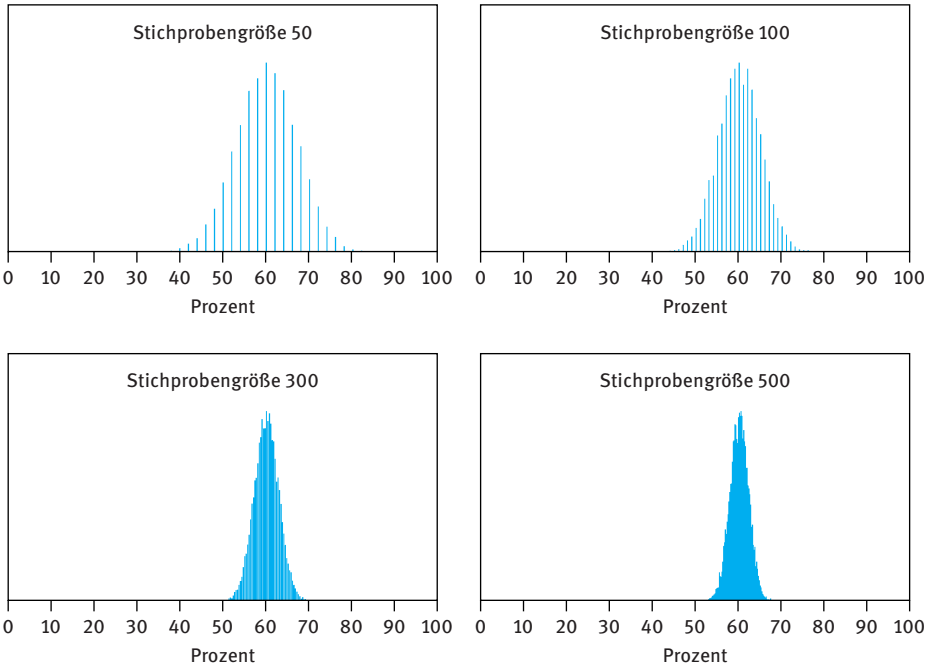

Abb. 2.19: Stichprobenvariabilität in Abhängigkeit von der Stichprobengröße (50, 100, 300 und 500 Personen) für den Prozentsatz an verheirateten Personen bei einem wahren Prozentsatz von 60 % in der Gesamtpopulation. Resultate von Computersimulationen mit jeweils 10.000 Stichproben.

liegen dabei zentriert um den wahren Wert herum. Die „Breite" der Normalverteilung muss allerdings geeignet gewählt werden. Eine Abbildung in Kap. 2.3 auf unserer Lehrbuch-Homepage zeigt geeignete Normalverteilungskurven für die Berechnung des Prozentsatzes verheirateter Personen (oben) sowie für die Berechnung des mittleren BMI-Wertes (unten), jeweils in Abhängigkeit von der Stichprobengröße. Diese entsprechen annähernd den Simulationsverteilungen, die in den oberen Hälften von Abb. 2.18 und auf einer Abbildung in Kap. 2.3 auf unserer Lehrbuch-Homepage zu sehen sind.

2.3.5 Die Normalverteilung in aller Kürze

Es ist sinnvoll, sich eingehender mit der *Normalverteilung* (Gauß-Verteilung) auseinander zu setzen. „Normal" bedeutet hier, dass diese statistische Verteilung in vielen Situationen einer guten Annäherung an die wahren Werte entspricht, sodass sie auch als Grundlage für Berechnungen dienen kann. Abb. 2.20 zeigt die *Dichtefunktion der Standard-Normalverteilung*. Die Dichtefunktion kann man sich als „geglättetes Histogramm" von unendlich vielen Werten vorstellen. Es stellt aber nicht die Anzahl der

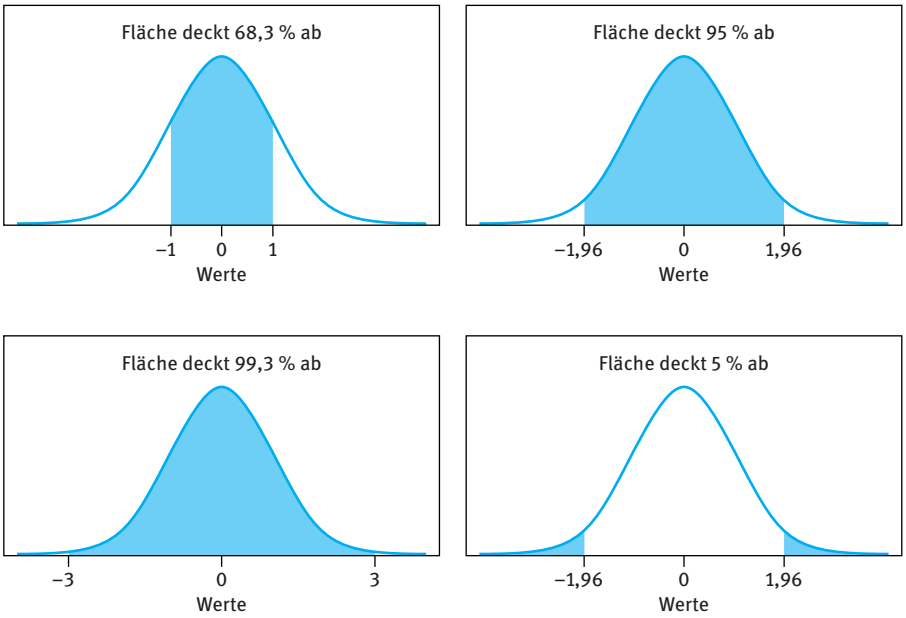

Abb. 2.20: Die Standard-Normalverteilung mit dem Mittelwert (MW) = 0 und der Standardabweichung (SD) = 1. Die gesamte Fläche zwischen der Linie der Dichtefunktion und der X-Achse beträgt 100 %.

beobachteten Werte dar, sondern die prozentuale Verteilung dieser Werte. Hierbei beträgt die gesamte Fläche zwischen der Funktionslinie und der X-Achse genau 100 %. Gibt man ein bestimmtes Intervall vor, dann ergibt die *Fläche unter der Kurve (engl. Area under the Curve)* den Prozentsatz der Werte in diesem Intervall. Im Intervall zwischen – 1 und + 1 befinden sich z. B. 68,3 % aller Werte. Da die Kurve symmetrisch zum Wert 0 ist, kann man daraus ableiten, dass 34,15 % der Werte zwischen 0 und + 1 liegen. Dem entsprechend befinden sich 95 % aller Werte zwischen – 1,96 und + 1,96. Fünf Prozent der Werte liegen damit weiter als 1,96 von 0 entfernt.

Leider ist es nicht möglich, eine einfache Formel anzugeben, mit der man diese Flächenabschnitte für jedes Intervall selber berechnen kann. Früher gab es daher in Statistikbüchern eine Tabelle, die die Werte für die Flächenabschnitte von minus Unendlich bis zu einem bestimmten Wert („*Z*-Wert“ genannt) angab. Heute kann man diese mit Hilfe verschiedener Computerprogramme berechnen (s. Internet-Ressourcen).

2.3.6 Das 95 %-Vertrauensintervall

Wir wissen nun, dass bei einer großen Stichprobe der hierbei berechnete Mittelwert bzw. ein bestimmter Prozentsatz der Variablen annähernd einer Normalverteilung um den wahren Wert in der Gesamtpopulation folgt. Damit ist es uns möglich, rund um den in der Stichprobe berechneten Wert (Mittelwert oder Prozentsatz) ein Intervall zu berechnen, das mit einer gewünschten Wahrscheinlichkeit den wahren Wert enthält. Was uns noch fehlt, ist eine Formel, die angibt, welche Breite die zu benützende Normalverteilung haben soll. Sie kann durch die Formeln für den so genannten Standardfehler berechnet werden. In Englisch wird der Standardfehler als *Standard Error* bezeichnet und mit „SE“ abgekürzt.

Den Standardfehler für einen Mittelwert errechnet man, indem man die Standardabweichung aller Werte in der Population durch die Quadratwurzel der Stichprobengröße dividiert.

Formel 2.2: Formeln für die Berechnung des Standardfehlers (a) eines Mittelwertes und (b) einer Proportion. SE = Standard Error (engl.).

$$\text{a. Standardfehler (SE) für Mittelwert} = \frac{\text{Standardabweichung der Werte}}{\sqrt{N}} = \frac{SD}{\sqrt{N}}$$

$$\text{b. Standardfehler (SE) für Proportion} = \sqrt{\frac{\text{Proportion} \times (1 - \text{Proportion})}{N}}$$

In der Praxis ist die Standardabweichung für alle Werte einer Population in der Regel jedoch nicht bekannt. Deshalb wird hierzu die Standardabweichung der vorliegenden Werte benützt. Beide Werte stimmen näherungsweise überein. Auch hier gilt: Die Annäherung ist umso besser, je größer die Stichprobe ist. Eine analoge Formel gibt es für den Standardfehler einer Proportion.

Um nun das **95 %-Vertrauensintervall** (**VI**, auch: Konfidenzintervall) zu berechnen, wird anschließend zum erhaltenen Mittelwert noch 1,96-mal der Standardfehler für diesen Wert auf der einen Seite addiert, auf der anderen Seite subtrahiert. Das gewählte Vertrauensintervall umfasst denjenigen Bereich um den geschätzten Wert herum, der mit einer zuvor festgelegten Wahrscheinlichkeit (hier: 95 %) die wahre Lage dieses Wertes angibt.

Formel 2.3: Formeln für die Berechnung des 95 %-Vertrauensintervalls eines Mittelwertes. MW = Mittelwert, SE = Standardfehler

$$\begin{aligned} \text{95 \%-Vertrauensintervall (VI) für Mittelwert} &= [MW - 1{,}96 \times SE(MW);\ MW + 1{,}96 \times SE(MW)] \\ &= MW \pm 1{,}96 \times SE(MW) \end{aligned}$$

Analog hierzu lässt sich auch das 95 %-Vertrauensintervall für eine Proportion berechnen.

Wenn nun z. B. bei einer randomisierten Behandlungsstudie zwei Gruppen miteinander verglichen werden, dann interessiert uns in der Regel eine Größe, die den Unterschied zwischen den beiden Gruppen beschreibt. Dies kann z. B. die Differenz zwischen den beiden Mittelwerten der betrachteten Gruppen sein oder die Differenz zwischen zwei Proportionen dieser Gruppen. Auch die Risiken, dass ein bestimmtes Ereignis wie Rückfall, Herzinfarkt oder Tod eintritt, können in beiden Gruppen unterschiedlich sein. Hier lässt sich ebenfalls ein 95 %-Vertrauensintervall analog zum oben beschriebenen Vorgehen konstruieren. Zum beobachteten Wert für die Differenz wird auf der einen Seite der Standardfehler für die interessierende Größe 1,96-mal addiert, auf der anderen Seite subtrahiert.

Entsprechende Formeln kommen bei der Berechnung des 95 %-Vertrauensintervalls eines *relativen Risikos* und der *Odds Ratio* zur Anwendung (s. a. Kap. 2.1). Hier muss jedoch beachtet werden, dass die Werte vor der Berechnung logarithmiert und dann zum Schluss auf die gewünschte Skala zurück transformiert werden müssen. Verwendet wird dabei der natürliche Logarithmus zur Basis der Eulerschen Zahl *e*. Die Rücktransformation muss daher mit der Exponentialfunktion e^x geschehen.

Es stellt sich nun natürlich die Frage, wie gut diese Annäherungen (*Approximationen*) in der praktischen Anwendung sind. Bei einer Anzahl von weniger als 60 Messwerten wird das 95 %-Vertrauensintervall für den Mittelwert zu eng, wenn man es mit dem Faktor 1,96 aus der Normalverteilung berechnet. In dieser Situation empfiehlt es sich, die Familie der t-Verteilungen heranzuziehen (s. dazu http://de.wikipedia.org/wiki/Studentsche_t-Verteilung), um den passenden Faktor für die Berechnung des 95 %-Vertrauensintervalls zu ermitteln. Bei einer Wertereihe von nur 50 (bzw. 20, 10 oder 5) Messungen sollte anstatt des Faktors 1,96 besser der Faktor 2 (bzw. 2,1; 2,23; 2,57) verwendet werden. Bei Proportionen, wie z. B. bei einem Risiko (= Anzahl Ereignisse / Anzahl Personen in der Gruppe), wird die approximative Berechnung des 95 % Vertrauensintervalls unzuverlässig, sobald die Anzahl der Ereignisse bzw. der Nichtereignisse kleiner als 5 wird. In diesen Fällen müssen anstatt der Annäherung über die Normalverteilung andere Methoden verwendet werden.

2.3.7 Der Umgang mit Wahrscheinlichkeiten: Interpretation von Untersuchungen und Tests

Obwohl die Wahrscheinlichkeitsrechnung als Teilgebiet der Mathematik nur wenige Rechenregeln kennt und daher auf den ersten Blick einfach erscheint, macht der Umgang mit Wahrscheinlichkeiten nicht selten Probleme. Eine Box in Kap. 2.3 auf unserer Lehrbuch-Homepage fasst wichtige Regeln der Wahrscheinlichkeitsrechnung zusammen.

Probleme entstehen besonders häufig bei der Interpretation der Resultate von Untersuchungen und Tests. Hier sind mehrere, unterschiedlich definierte *bedingte Wahrscheinlichkeiten* von Bedeutung. So ist die **Sensitivität** eines Tests die Wahrscheinlichkeit, dass der Test bei einer tatsächlich erkrankten Person positiv ausfällt. In mathematischer Schreibweise wird dies folgendermaßen ausgedrückt: P(positiver Test | Person hat die Krankheit). Als **Spezifität** eines Tests bezeichnet man dagegen die Wahrscheinlichkeit, dass bei einer nicht erkrankten Person der Test auch tatsächlich negativ ausfällt. Die Notation lautet hier: P(negativer Test | Person hat die Krankheit nicht). Es handelt sich hierbei um bedingte Wahrscheinlichkeiten, da man jeweils das Eintreten eines Ereignisses *A* (positiver/negativer Test) unter der Bedingung anschaut, dass ein anderes Ereignis *B* (Person hat eine Krankheit/keine Krankheit) eingetreten ist.

Was ist nun die Wahrscheinlichkeit für einen falsch positiven Test? Um dies zu beantworten gilt es, die Frage zu präzisieren. Bezieht sich diese Aussage auf nicht erkrankte Personen, dann ist die Wahrscheinlichkeit für einen falsch positiven Test einfach gleich 1 – Spezifität: Die beiden Wahrscheinlichkeiten verhalten sich komplementär (s. dazu eine Box in Kap. 2.3 auf unserer Lehrbuch-Homepage). In der Praxis bezieht sich die Frage jedoch häufig auf die Personen mit positiven Testresultaten. Um diese beraten zu können, muss man wissen, wie häufig Personen mit einem positiven Test die Krankheit nicht haben: P(Person hat die Krankheit nicht | positiver Test). Wir können dies am Beispiel von systematisch durchgeführten Mammografien bei 50-jährigen Frauen zur Früherkennung von Brustkrebs illustrieren. Um die Frage beantworten zu können, benötigen wir einige zusätzliche Angaben:

- Die *Sensitivität* der Mammografie-Untersuchung liegt bei 85 %, d. h. bei 100 Frauen mit Brustkrebs wird der Test in 85 Fällen positiv ausfallen.
- Die *Spezifität* der Mammografie-Untersuchung liegt bei 97 %, d. h. bei 100 Frauen ohne Brustkrebs wird der Test in 3 Fällen positiv ausfallen.
- Weiter wird angenommen, dass von tausend 50-jährigen Frauen, die sich gesund fühlen, zwei unerkannt an Brustkrebs erkrankt sind (*Prävalenz*, s. a. Kap. 2.1).

Diese realistischen Annahmen liegen den Berechnungen in Tab. 2.8 zugrunde. Hier wurden 10.000 Frauen mammografiert. Da nach unserer Annahme zwei von 1.000 sich gesund fühlenden Frauen an Krebs erkrankt sind, haben in unserer Gruppe 20 Frauen Brustkrebs. Die restlichen 9.980 Frauen sind nicht an Brustkrebs erkrankt. Bei einer Test-Sensitivität von 85 % hätten 17 der 20 Frauen mit Brustkrebs ein positives Test-Resultat. Drei Brustkrebsfälle blieben dagegen unerkannt. Darüber hinaus gäbe es auch bei 3 % (= 299 Frauen) der 9.980 Frauen ohne Brustkrebs ein positives Test-Resultat (Spezifität 97 %). Aus diesen Berechnungen ergibt sich, dass insgesamt 299 der 316 (17 + 299) positiven Tests falsch positiv sind, was 94,6 % entspricht!

Tab. 2.8: Interpretation der Resultate eines Tests am Beispiel eines Mammografie-Screenings bei 10.000 Frauen.

Annahmen:
- Die Sensitivität des Tests ist 85 %
- Die Spezifität des Tests ist 97 %.
- Die Prävalenz der Krankheit beträgt 2 von 1.000 (also 20 von 10.000)

Testresultat	Personen mit der Krankheit	Personen ohne die Krankheit	Gesamt
Positiv	17	299	316
Negativ	3	9.681	9.684
Gesamt	20	9.980	10.000

- Der *positiv prädiktive Wert* des Tests beträgt 17/316 = 5,4 %.
- 94,6 % der positiven Tests sind falsch positiv.
- Der *negativ prädiktive Wert* des Tests beträgt 9.681/9.684 = 99,97 %.

Als **positiv prädiktiven Wert** (*PPV = Positive Predictive Value*) bezeichnet man den Anteil der tatsächlich erkrankten Personen unter allen Personen mit positivem Test: P(Person hat die Krankheit | positiver Test). In der geschilderten Situation wären dies 17 von 317 Frauen, d. h. nur 5,4 % der Frauen mit positivem Test wären tatsächlich erkrankt. Führt man dieselbe Berechnung mit einem anderen Prävalenzwert durch, ändert sich auch der PPV. Je höher die Krankheitshäufigkeit ist, desto höher liegt auch die Zahl der tatsächlich Erkrankten unter den positiv getesteten Personen und damit der PPV. Dies ist einer der Gründe, warum das Mammografie-Screening bei Frauen unter 50 Jahren nicht empfohlen wird: Brustkrebs ist in dieser Altersgruppe weniger häufig als bei älteren Frauen.

Der **negative prädiktive Wert** (*NPV = Negative Predictive Value*) ist definiert als der Anteil der tatsächlich gesunden Personen unter allen Personen mit negativem Test: P(Person hat die Krankheit nicht | negativer Test). In unserem Fall wären das 9.681 von 9.684 Frauen. Dies bedeutet, dass 99,97 % aller Frauen mit einem negativen Testergebnis tatsächlich nicht an Brustkrebs erkrankt waren. Hier gilt: Je niedriger die Prävalenz, desto höher ist der NPV. Um sich diese Zusammenhänge einzuprägen, wiederholen Sie am besten die Berechnungen in Tab. 2.7 mit einer Prävalenz von 20 %.

Eine ausführliche Diskussion über die Vor- und Nachteile von Screening-Untersuchungen finden Sie in Kap. 4.5.

2.3.8 Statistische Signifikanz und p-Wert

Oft liest man in wissenschaftlichen Zeitschriften, dass die Resultate einer Studie „statistisch signifikant" seien. Um zu erläutern, was damit gemeint ist, betrachten wir die Resultate einer im Jahr 2010 im englischen Medizinjournal *The Lancet* veröffentlichten randomisierten Studie. Die Studie untersuchte, ob eine einmalige Sigmoidoskopie (= endoskopische Untersuchung des Enddarms einschließlich der S-förmigen Grimmdarmschlinge) bei klinisch gesunden Personen im Alter von 55 bis 64 Jahren die Darmkrebs-Sterblichkeit in den nächsten 11 Jahre reduziert. Als Vergleichsgruppe dienten Gleichaltrige, bei denen keine solche Untersuchung durchgeführt wurde. Die Studienautoren untersuchten neben der Darmkrebs-Sterblichkeit auch die Gesamtsterblichkeit in beiden Gruppen (Tab. 2.9). Die Berechnungen ergaben, dass die Darmkrebs-Sterblichkeit in der Sigmoidoskopie-Gruppe im Vergleich zu Kontrollgruppe um 31 % gesenkt werden konnte. Das relative Risiko (RR) betrug 0,69 bei einem 95 % VI von 0,59 bis 0,80. Die Gesamtsterblichkeit sank dadurch nach Angaben der Autoren um 3 % (RR: 0,97; 95 %-VI: 0,95; 1,00).

Die Autoren veröffentlichten zusätzlich den so genannten **p-Wert**, der in der englischen Terminologie als „p-value" bezeichnet wird. Auch der p-Wert ist eine *bedingte* Wahrscheinlichkeit. Für Studien, die die Wirksamkeit einer bestimmten Intervention untersuchen, wird in der Regel als Bedingung die so genannte „**Null-Hypothese**" gewählt. Bei dieser Hypothese geht man davon aus, dass die Intervention keine Wirkung hat. Unter dieser Annahme wird nun die Wahrscheinlichkeit berechnet, dass der tatsächlich beobachtete oder ein noch größerer Unterschied rein zufällig zustande gekommen sind.

Bei der Sigmoidoskopie-Studie sagt der p-Wert von < 0,0001 für die Darmkrebs-Sterblichkeit folgendes aus: Unter der Annahme, dass eine einmalige Sigmoidoskopie die Darmkrebs-Sterblichkeit bei den untersuchten Personen nicht reduziert – das relative Risiko also 1 ist –, ist die Wahrscheinlichkeit kleiner als 1 zu Zehntausend, ein relatives Risiko von ≤ 0,69 oder ≥ 1,45 (= 1/0,69) rein zufällig zu beobachten. Bei der Analyse der Gesamtsterblichkeit nennen die Autoren einen p-Wert von 0,052. Dies bedeutet analog, dass unter der Annahme, die einmalige Sigmoidoskopie reduziere die Gesamtsterblichkeit bei den untersuchten Personen nicht, eine Wahrscheinlichkeit von 5,2 % besteht, ein relatives Risiko von ≤ 0,97 oder ≥ 1,03 (1/0,97) zu beobachten, das rein durch Zufall zustande gekommen ist. Diese p-Werte werden „zweiseitig" genannt, weil hier die Entfernung zum Null-Wert, der für „keine Wirksamkeit" steht, sowohl nach oben („Nutzen durch Behandlung") als auch nach unten („Schaden durch Behandlung") betrachtet wird.

Tab. 2.9: Randomisierte Studie zur Wirksamkeit einer einmaligen Sigmoidoskopie als Mittel der Darmkrebs-Früherkennung: Zahl der Todesfälle insgesamt sowie der Darmkrebs-Todesfälle in beiden Gruppen während eines Zeitraums von etwa 11 Jahren (Resultate übernommen aus Lancet 2010; 375: 1624–33, Tab. 1).

	Gruppe mit einmaliger Sigmoidoskopie	Kontrollgruppe ohne Sigmoidoskopie	Relatives Risiko (95 %-VI)	p-Wert
Darmkrebs-Todesfälle	221	637	0,69 (0,59; 0,80)	< 0,0001
Alle Todesfälle	6.775	13.768	0,97 (0,95; 1,00)	0,052
Gesamtzahl der Personen	57.099	112.939		

95 %-VI: 95 %-Vertrauensintervall

Das *relative Risiko* (RR) vergleicht die Gruppe, bei deren Mitgliedern jeweils eine einmalige Sigmoidoskopie durchgeführt wurde, mit der Kontrollgruppe (s. Kap. 2.1.3).

Die Berechnung des p-Wertes

Der **p-Wert** lässt sich unter Verwendung der *Standard-Normalverteilung* in drei Schritten berechnen.

- Zuerst berechnet man die Distanz zwischen dem Studien-Wert für die Wirksamkeit einer Methode und der Null-Hypothese, d. h. dem Wert, der „keine Wirksamkeit" beschreibt. Wenn *relative Risiken* (RR) betrachtet werden, müssen die Berechnungen auf der logarithmischen Skala durchgeführt werden. Für die Gesamtsterblichkeit bei der betrachteten Sigmoidoskopie-Studie berechnet sich dies aus ln(0,97332) – ln(1). Als Ergebnis erhalten wir – 0,02704.
- Der Absolutbetrag dieses Resultates wird anschließend durch den Standardfehler dividiert. Auch hier muss bei relativen Risiken die logarithmische Skala verwendet werden. Berechnet man den Logarithmus des Standardfehlers SE(ln(RR)), so ergibt dies 0,01392. Teilt man nun 0,02704 durch 0,01392, erhält man den Wert 1,942529. Dieser Wert wird als **Z-Wert** zur Berechnung des p-Wertes bezeichnet.
- In einem dritten Schritt wird nun berechnet, welcher Prozentsatz der Werte bei der Standard-Normalverteilung weiter als Z von Null entfernt liegt. In unserem Beispiel bedeutet dies: Welcher Prozentsatz ist kleiner/größer als der errechnete Z-Wert, d. h. kleiner als – 1,942529 oder größer als 1,942529? Wir wissen, dass bei der Standard-Normalverteilung genau 5 % aller Werte außerhalb von ±1,96 liegen. Also erwarten wir etwas mehr als 5 %. Die genaue Berechnung ergibt 5,2 %.

Auch für die Differenz der Mittelwerte aus zwei Behandlungsgruppen lässt sich analog ein p-Wert berechnen. In die Berechnung des p-Wertes fließt also der *Standardfehler* und dadurch auch die *Größe der Studie* mit ein.

Dualität zwischen 95 %-Vertrauensintervall und „statistischer Signifikanz“

Es hat sich eingebürgert, dass p-Werte, die kleiner als 0,05 sind, als „statistisch signifikant“ bezeichnet werden. Die Wahl der 0,05-Grenze hat den Vorteil, dass eine Dualität zwischen dem 95 %-Vertrauensintervall und der „statistischen Signifikanz“ besteht. In den Fällen, in denen das 95 %-Vertrauensintervall den Wert für „keine Wirksamkeit“ ausschließt, ist der p-Wert kleiner als 0,05 und damit das Resultat „statistisch signifikant“ (und umgekehrt).

Dies sehen wir z. B. bei den Darmkrebstodesfällen in der Sigmoidoskopie-Studie. Das 95 %-Vertrauensintervall für das relative Risiko reicht von 0,59 bis 0,80 und schließt damit den Wert 1 (= „keine Wirksamkeit“) klar aus. Entsprechend ist der p-Wert deutlich kleiner als 0,05. Dagegen reicht das 95 %-Vertrauensintervall für das relative Risiko bei der Gesamtsterblichkeit von 0,95 bis 1,00. Es berührt also den Wert 1, der für „keine Wirksamkeit“ steht. Aufgrund der Dualität zwischen dem 95 %-Vertrauensintervall und „statistischer Signifikanz“ sollte hier der p-Wert bei 0,05 (= 5 %) liegen. Gibt man das Resultat mit mehr als zwei Stellen nach dem Komma an, sieht man, dass das obere Ende des 95 %-Vertrauensintervalls 1,00024 beträgt. Es schließt also die 1 noch knapp mit ein. Damit muss der p-Wert etwas größer als 5 % sein.

2.3.9 Statistische Signifikanz und klinische Relevanz

Statistische signifikante Resultate sind nicht zwingend auch klinisch relevant. In Tab. 2.10 sind die hypothetischen Resultate von drei randomisierten placebokontrollierten Studien zur Senkung des LDL-Cholesterins im Blut dargestellt. In allen drei Studien wurden die TeilnehmerInnen zufällig entweder derjenigen Gruppe zugeteilt, in der sie das neue Medikament (A, B oder C) erhielten oder der Placebo-Gruppe. Dort wurde ihnen statt des zu testenden Medikaments ein Scheinmedikament (Placebo) verabreicht. Nach einer Behandlungsdauer von 3 Monaten wurde in allen Gruppen der Blutspiegel des LDL-Cholesterins gemessen. Daraus wurden nun die Mittelwerte pro Behandlungsgruppe sowie die Differenzen der Mittelwerte berechnet (Spalte 4 von Tab. 2.10). Anschließend wurde der Standardfehler für die Differenz von Mittelwerten (Spalte 5) und die Grenzen des 95 %-Vertrauensintervalls (Spalte 6) ermittelt. Zum Schluss wurden der Z-Wert sowie der p-Wert berechnet (Spalten 7 und 8).

Interessanterweise ergaben die Berechnungen sowohl für Medikament A als auch für Medikament B den gleichen p-Wert von 0,544. Die Resultate sind also beide statistisch *nicht signifikant*. Das erstaunt nicht. In beiden Studien ist zu sehen, dass der Wert 0 deutlich im 95 %-Vertrauensintervall enthalten ist. Betrachtet man die Grenzen des 95 % Vertrauensintervalls von Studie 1, fällt auf, dass der Behandlungseffekt hiermit nicht sinnvoll eingegrenzt wurde. Er liegt mit 95 % Wahrscheinlichkeit zwischen einer Senkung um 84,7 mg/dl und einer Erhöhung um 44,7 mg/dl. Da in Studie 1 nur 40 Patienten pro Gruppe untersucht wurden, überrascht dieses

unpräzise Resultat nicht. Es sind also größere Studien notwendig, um die Wirksamkeit von Medikament A abzuklären. In Studie 2 umfasste jede Gruppe 4.000 Personen. Die Wirksamkeit von Medikament B konnte recht präzise quantifiziert werden. Sie liegt mit 95 % Wahrscheinlichkeit zwischen einer Senkung um 8,47 mg/dl und einer Erhöhung um 4,47 mg/dl. Eine klinisch relevante Senkung um ≥10 mg/dl ist daher sehr unwahrscheinlich. Bei Medikament C liegt mit einen p-Wert von 0,012 ein *statistisch signifikanter* Behandlungseffekt vor. Der Wert 0 ist nicht im 95 %-Vertrauensintervall enthalten. Betrachtet man die Grenzen des 95 %-Vertrauensintervalls, so liegt der Behandlungseffekt mit 95 % Wahrscheinlichkeit zwischen einer Senkung um 8,92 mg/dl und einer Senkung um 1,08 mg/dl. Damit liegt zwar eine „statistisch signifikante" Senkung vor, aber auch hier ist eine Senkung um ≥10 mg/dl eher unwahrscheinlich.

Um alle drei Studien abschließend beurteilen zu können, ist es wichtig zu wissen, welches Ausmaß einer Senkung des LDL-Cholesterinspiegels im Blut klinisch relevant ist. Geht man davon aus, dass dies erst bei einer Senkung um mindestens 10 mg/dl der Fall ist, dann ist auch Medikament C nicht geeignet, da es ja nur mit einer sehr geringen Wahrscheinlichkeit eine solche Senkung erreicht. Es zeigt sich, dass die Information, ob ein Behandlungseffekt statistisch signifikant ist, allein nicht ausreicht, um die klinische Relevanz der Resultate einer Studie beurteilen zu können. Die Information des 95 %-Vertrauensintervalls ist hier wesentlich nützlicher. Wir erhalten einen 95 %-Wahrscheinlichkeitsbereich für den Behandlungseffekt und können daraus auch ableiten, ob der Behandlungseffekt in einem Bereich liegt, der klinisch relevant ist.

Tab. 2.10: Hypothetische Resultate von drei placebokontrollierten, randomisierten Studien zur Senkung des LDL-Cholesterins im Blut.

Studie	Medikament	Anzahl der Patienten pro Gruppe	Differenz der Mittelwerte des LDL-Cholesterins (mg/dl) zwischen der Medikamenten-Gruppe und der Placebo-Gruppe	Standardfehler für die Differenz der Mittelwerte des LDL-Cholesterins	Grenzen des 95 %-Vertrauensintervalls für die Differenz der Mittelwerte des LDL-Cholesterins	Z-Wert	p-Wert
1	A	40	−20	33	−84,68 bis 44,68	−0,606	0,544
2	B	4.000	−2	3,3	−8,47 bis 4,47	−0,606	0,544
3	C	5.000	−5	2	−8,92 bis −1,08	−2,5	0,012

2.4 Sozialwissenschaftliche Datenerhebung

Siegfried Geyer, Thomas Abel

Anders als in der Medizin werden die für Forschung und Praxis nötigen Daten im Bereich der *Sozialwissenschaften* in erster Linie über Fragebögen und nicht durch die Messung biologisch-medizinischer Parameter gewonnen. In der Gesundheitsförderung und der Prävention setzt man Fragebögen häufig dann ein, wenn man etwas über das Wissen, die Wahrnehmungen oder subjektiven Beurteilungen zu bestimmten Verhaltensweisen, Zuständen oder Bedürfnissen von Personen bzw. Personengruppen erfahren möchte. Solche systematischen Befragungen, die das Ziel haben, Daten zu einem bestimmten Thema zu erheben, nennt man auch *Surveys*. Kenntnisse in der Entwicklung und Anwendung von Fragebögen sind unentbehrlich, wenn es darum geht, Public-Health-Studien zu beurteilen oder gar selbst durchzuführen.

2.4.1 Was ist eine gute Frage?

Eine klare und verständliche Formulierung der Fragen ist die wichtigste Voraussetzung dafür, dass sich aus den mit Hilfe von Fragebögen erhobenen Daten später durch Interpretation auch Schlüsse ziehen lassen. Hierzu müssen ForscherInnen und Befragte eine gestellte Frage in gleicher Weise verstehen und interpretieren können. Dies ist in der Praxis keineswegs selbstverständlich. Denn nicht immer sprechen die Konstrukteure eines Fragebogens und die Adressaten, an die sich der Fragebogen richten soll, im Hinblick auf den Wortschatz und das sprachliche Niveau die gleiche Sprache. Wenn eine Frage verstanden wurde, dann müssen die Befragten auch über die notwendige Information verfügen, sie zu beantworten. Dazu gehört nicht nur das Wissen um eine Antwort, sondern auch genügend Zeit, um sich an die Information zu erinnern und die Antwort dann zu formulieren.

In Lebensqualitätsfragebögen wird z. B. danach gefragt, wie häufig bestimmte Symptome innerhalb eines definierten Zeitraums aufgetreten sind. Da jedoch Ereignisse, die als wenig relevant erachtet wurden, aufgrund der Struktur des menschlichen Gedächtnisses nach einer gewissen Zeit vergessen werden, sind die Antworten hierauf unter Umständen wenig präzise. Seltene Ereignisse, wie etwa die Häufigkeit des Auftretens von Symptomen oder die Zahl von Arztbesuchen, werden in der Regel gezählt. Bei häufigeren Ereignissen basieren die Angaben dagegen auf groben Schätzungen und sind entsprechend ungenau.

Antworten werden in der Regel durch solche Sachverhalte bestimmt, die zum Zeitpunkt der Fragestellung im Gedächtnis der Befragten präsent sind. Ist ein längerer Erinnerungsprozess erforderlich, muss den Befragten genügend Zeit zur Verfügung stehen. Jedoch auch dann können im Ergebnis erhebliche Urteilsfehler auftreten. So kann z. B. die Zahl der Arztbesuche falsch eingeschätzt werden, wenn sich die Befragten an ein bestimmtes Datum nicht direkt erinnern können. Oft wird es dann

aus anderen Ereignissen rekonstruiert. Bei dieser Rekonstruktion können jedoch Irrtümer vorkommen. Schließlich können Fragen, die den Befragten peinlich oder in anderer Weise unangenehm sind, zu einer Antwortverweigerung führen. Beispiele hierfür sind Fragen nach dem Alkoholkonsum, nach Sexualpraktiken oder auch nach dem Einkommen.

Die Qualität der gegebenen Antworten ist jedoch nicht nur von der Verständlichkeit der Fragen abhängig, sondern auch von der Länge des Fragebogens. Mit zunehmender Befragungsdauer nehmen Konzentrationsprobleme bei den Befragten zu, das Risiko von Urteilsfehlern steigt, während die Motivation zur Teilnahme sinkt. Dies ist insbesondere bei alten Menschen und Menschen mit Erkrankungen zu berücksichtigen.

Bei der Konstruktion eines Fragebogens ist die Entscheidung, ob die Antwortmöglichkeiten vorgegeben (sog. *geschlossene Fragen*) oder die Antworten offen gelassen werden (sog. *offene Fragen*), vom Verwendungszweck und der geplanten Vorgehensweise bei der Auswertung abhängig. Fragen mit vorgegebenen Antwortmöglichkeiten sind in der Regel schneller zu beantworten, die quantitativen Informationen sind leichter auszuwerten. Geschlossene Fragen grenzen jedoch den Antworthorizont der Befragten auf die vorgegebenen Alternativen ein, und zwar auch dann, wenn die zusätzliche Option einer offenen Antwort vorgegeben wird. Die Antwortvorgaben bei geschlossenen Fragen sollten immer einen möglichst hohen Grad an Eindeutigkeit haben.

Beispiel für eine geschlossene Frage:

Wie häufig haben Sie in den letzten sechs Monaten wegen einer Erkrankung oder wegen Beschwerden eine Arztpraxis aufgesucht?

Antwortmöglichkeiten:
○ gar nicht ○ einmal ○ zwei- bis viermal ○ mehr als viermal

Wenn über den Gegenstand einer Frage wenig bekannt ist, sollten die Antworten offen gelassen werden. Die Antworten auf solche offenen Fragen sind meist subjektive Einschätzungen der Befragten, in die eine möglichst große Bandbreite an Informationen einfließen sollte. Offene Fragen liefern v. a. qualitative Informationen. Ihre Auswertung ist meist aufwendig.

Beispiel für eine offene Frage:

Gibt es Ihrer Meinung nach Zusammenhänge zwischen Ihrer Arbeitslosigkeit und Ihrem Gesundheitszustand? Und wenn ja, welche?

Antwort: (Bitte verwenden Sie so viele Zeilen, wie Sie möchten.)

2.4.2 Was führt zu einer guten Antwort?

Bei der Konstruktion von Fragebögen kann man auf mehrere Antwortformat-Optionen zurückgreifen. Je nach Verwendungszweck können sie innerhalb eines Fragebogens auch miteinander kombiniert werden. Eine Abbildung in Kap. 2.4 auf unserer Lehrbuch-Homepage zeigt einen Abschnitt aus einem Fragebogen der Eidgenössischen Jugendbefragung *CH-X 2010 – Vertiefungsfragen zur Gesundheit*, bei dem verschiedene Antwortformat-Optionen verwendet wurden.

Ratingskalen/Ordinalskalen

Am häufigsten werden Ratingskalen verwendet, die mehrere Antwortalternativen anbieten und den Befragten dadurch eine abgestufte Antwort ermöglichen. Dabei sollten die Alternativen so formuliert werden, dass sich die einzelnen Kategorien auf den gleichen Inhalt beziehen und semantisch die gleichen Abstände haben. Diese *semantische Äquidistanz* wurde bisher für drei Beurteilungsdimensionen untersucht:

- Häufigkeit: nie – selten – gelegentlich – oft – immer
- Intensität: nicht – wenig – mittelmäßig – ziemlich – sehr
- Bewertung von Aussagen: stimmt nicht – stimmt wenig – stimmt mittelmäßig – stimmt ziemlich – stimmt sehr

Über die bestmögliche Gestaltung von Antwortformaten gab es lange Zeit Unklarheit. In den letzten Jahren wurden jedoch Studien durchgeführt, die es erlauben, einige der strittigen Fragen zu beantworten. Daraus können die folgenden Empfehlungen abgeleitet werden:

- Die Zahl der Antwortstufen sollte unter Berücksichtigung der zu beurteilenden Thematik und der sprachlichen und intellektuellen Fähigkeit der Befragten innerhalb eines Bereichs von 5±2 Kategorien bleiben. Für Standardanwendungen sind Fünfpunktskalen eine gute Lösung.
- Skalen mit weniger als 5 Punkten sind wenig reliabel (*Reliabilität* s. Kap. 2.1.4), da die Befragten bei der Wahl der zutreffenden Kategorie häufiger unsicher sind.
- Vierpunktskalen, die deshalb eingesetzt werden, weil man Mittelkategorien vermeiden möchte, werden ihr Ziel verfehlen. Sie werden mit großer Wahrscheinlichkeit verstärkt zur Wahl von extremen Antworten führen.
- In Antwortskalen sollte jeder Skalenpunkt verbal bezeichnet sein. Dazu sind semantisch gleichabständige Begriffe (s. o.) zu verwenden.

Kategorialskalen

Eine solche Skala besteht aus sich gegenseitig ausschließenden Kategorien, die qualitativer Art und ohne eine natürliche Ordnung sind. Ein Beispiel hierfür ist die Klassifizierung von Personen nach ihrem Familienstand (schweizerisch: Zivilstand) in die Kategorien „ledig“, „verheiratet“, „geschieden“ oder „verwitwet“. Die Befrag-

ten können dort in Abhängigkeit von der Instruktion entweder nur eine oder auch mehrere Antworten ankreuzen. Mehrere Antworten könnten z. B. auch bei einer Frage nach der Art von vorhandenen Stress-Symptomen ausgewählt werden. In anderen Fällen können die Befragten aufgefordert werden, Begriffe in eine Rangreihe zu bringen.

Die Testung von Fragebögen

Es ist nun keineswegs sicher, dass ein Fragebogen in der Form, wie er entwickelt wurde, ohne weiteres auch später in der Praxis verwendet werden kann. Da Surveyfragen meist von Fachleuten entworfen werden, muss die verwendete Sprache nicht mit der der Zielgruppe übereinstimmen. In der praktischen Anwendung kann es zu Problemen kommen, wenn die Befragten eine Frage anders verstehen als von den Fragebogenkonstrukteuren gedacht. Auch können die verwendeten Begriffe mehrdeutig sein und dann von Befragten und Fragebogenkonstrukteuren unterschiedlich verstanden werden. Beides kann später zu erheblichen Schwierigkeiten in der Interpretation der gewonnenen Daten führen. Darüber hinaus können abstrakte Begriffe in ihrem inhaltlichen Verständnis divergieren. So kann z. B. eine Frage nach dem schweizerischen Gesundheitssystem so beantwortet werden, dass Befragte, die im Versicherungswesen arbeiten, bei ihrer Beantwortung primär das Versicherungssystem im Blick haben. ÄrztInnen denken dagegen in erster Linie an die ärztliche Versorgung. PatientInnen beantworten die Frage vor dem Hintergrund ihrer eigenen Erfahrung mit ÄrztInnen bzw. Einrichtungen der medizinischen Versorgung.

Bei der Lösung der daraus resultierenden Probleme können routinemäßig angewandte *Standardpretests* eine Hilfe sein. Hierbei werden die entwickelten Fragebögen in Interviews unter möglichst realistischen Befragungsbedingungen getestet. Die Interviewer registrieren dort die von den Befragten unaufgefordert abgegebenen Kommentare und melden diese an die Studienleitung zurück. Das Verfahren kann nur grobe Fehler aufdecken. Antworten von Befragten, die irrtümlich der Überzeugung sind, dass sie eine Frage korrekt verstanden haben, bleiben ungeprüft als richtig stehen. Nach dem derzeitigen Wissensstand können Standardpretests zur Schätzung des für ein Interview notwendigen Zeitaufwands dienen, nicht jedoch zur Aufdeckung von solch spezifischen Verständnisproblemen. Den Fragebogenkonstrukteuren steht mittlerweile ein umfangreiches Instrumentarium zur Testung der Verständlichkeit von Surveyfragen zur Verfügung. Nach dem derzeitigen Stand der Methodenforschung muss der Einsatz eines nicht getesteten Fragebogens als Fehler gewertet werden. Das am häufigsten verwendete Testverfahren ist das *Probing*. Hierbei werden potentiell unklare Begriffe oder auch eine ganze Frage auf ihre Verständlichkeit hin untersucht. Bislang gibt es noch keine komplette Liste von standardisierten Regeln zur Überprüfung von Fragebögen. Auch die angemessene Fallzahl für einen Pretest ist nicht festgelegt. Wenn jedoch komplexere Inhalte abgefragt werden und/oder die Grund-

gesamtheit der Befragten heterogen ist, werden größere Fallzahlen (ca. 20 Fälle) als angemessen erachtet.

2.4.3 Der Datenzugang über Surveys

Bei der Durchführung von Surveys wird zunächst entschieden, auf welche Weise die zu Befragenden ausgewählt werden sollen. Weiterhin wird festgelegt, ob ein Quer- oder ein Längsschnittdesign angewandt werden soll (s. Kap. 2.1.5). Die für eine Studie erforderliche *Fallzahl* richtet sich nach der Komplexität der Fragestellung, nach der Größe der zu untersuchenden Gruppen und nach der erwarteten Höhe der statistischen Effekte.

Wenn es Ziel einer Untersuchung ist, Aussagen über eine gesamte Bevölkerung zu machen, wird üblicherweise eine *Zufallsstichprobe* gezogen. Dabei können z. B. Daten von Einwohnermeldeämtern oder andere vollständigen Verzeichnissen genutzt werden. Ist die Zufallsstichprobe hieraus ausreichend groß, kann sie als verkleinertes Abbild der zu untersuchenden Bevölkerung angesehen werden und erlaubt die Verwendung statistischer Prüfverfahren. Die Ziehung von Zufallsstichproben ist jedoch nicht überall möglich, weshalb auf andere, manchmal auch auf suboptimale Verfahren ausgewichen werden muss.

Quotenstichproben werden gezogen, wenn die zu untersuchenden Gruppen unterschiedlich groß oder Subgruppen zu klein sind. Auch wenn eine Zufallsziehung aus verschiedenen Gründen nicht möglich ist, sind Quotenstichproben eine Alternative. Hierbei werden Quotenanteile von bestimmten Merkmalen der zu Befragenden (z. B. Geschlecht, Alter oder Berufsgruppen) vorgegeben. Die Details der Ziehung bleiben den Ausführenden der Befragung überlassen. Typische Probleme bei Quotenstichproben können unkontrollierbare Selektionseffekte (*Selektionsbias*, s. Kap. 2.1.8) sein, sodass Aussagen über die Grundgesamtheit aufgrund dieser systematischen Fehler nicht möglich sind.

Ein weiteres, nicht zufallsgesteuertes Verfahren ist das *Schneeballsystem*, das angewandt wird, wenn bei Gruppen relevante Merkmale nicht bekannt oder als Auswahlkriterien nicht verwendbar sind. Man beginnt bei einer Person der zu untersuchenden Gruppe. Von ihr ausgehend werden weitere Personen rekrutiert. Der Nachteil ist wiederum das unklare Verhältnis zur Grundgesamtheit. Selektionseffekte können hier selbst annäherungsweise nicht geschätzt werden.

Die Mehrzahl der bisher in der Gesundheitsforschung durchgeführten Studien basiert auf einem *Querschnittsdesign*, das nur eine Messung vorsieht. Da derartige Untersuchungen nur Momentaufnahmen ermöglichen, werden zunehmend häufiger *Längsschnittanalysen* durchgeführt. Sie ermöglichen die Untersuchung von Veränderungen, haben jedoch das Problem, dass über die Zeit ein Teil der Befragten die Studie verlässt. Auch dies kann zu Selektionseffekten (*Loss-to follow-up-Bias*) führen.

2.4.4 Standardisierte Methoden zur Erhebung von Daten

Persönliche Befragung

Die klassische Form der Befragung ist das persönliche Interview. Aus Kostengründen wird mittlerweile jedoch die Mehrzahl der Befragungsstudien mit Hilfe anderer Methoden durchgeführt. Bei der persönlichen Befragung sitzen sich Befragte und Interviewer gegenüber. Normalerweise verliest der/die InterviewerIn die Fragen, und die darauf gegebenen Antworten des/der Befragten werden registriert. Es ist auch möglich, Antwortalternativen in Form von Karten vorzulegen. Darüber hinaus können wahlweise Abbildungen, Modelle oder Fragebögen zum Selbstausfüllen eingesetzt werden.

Werden bei persönlichen Interviews Papierfragebögen eingesetzt, dann müssen die auf diese Weise gewonnenen Informationen anschließend in eine elektronische Form gebracht werden. Durch den Einsatz von Computern direkt bei der Befragung ist dies heute meist nicht mehr nötig. Solche „*Computer-Assisted Personal Interviews*" (CAPI) ermöglichen es, Kontrollen in die Dateneingabe einzubauen und nötige Korrekturen unmittelbar vornehmen zu lassen. Dabei wird durch ein Hintergrundprogramm automatisch geprüft, ob ein eingegebener Wert innerhalb eines definierten Bereichs liegt. Nach dem Abschluss der Befragung liegt dann bereits ein auswertungsfähiger Datensatz vor.

Durch die persönliche Form der Kommunikation kommt den InterviewerInnen bei dieser Form der Datenerhebung eine besondere Rolle zu. Sie müssen von den Befragten akzeptiert werden und – in Abhängigkeit von der Studienthematik – auch in der Lage sein, ein gewisses Vertrauensverhältnis aufzubauen. Hierzu ist eine gründliche Schulung der InterviewerInnen notwendig. Lange Zeit lernte man in solchen Schulungen, dass das Interviewer-Verhalten eher distanziert und auf die alleinige Gewinnung von Informationen ausgerichtet sein sollte. Untersuchungen haben jedoch gezeigt, dass diese Form von den Befragten oft als kalt und teilnahmslos empfunden wird. Eine emotional warme und unterstützende Form der Befragung erzielt bei inhaltlich neutraler Gesprächsführung deutlich bessere Daten. InterviewerInnen sollten dabei über ein ausreichendes Selbstbewusstsein verfügen und in der Lage sein, potentielle Befragte zu einer Teilnahme zu animieren. Darüber hinaus müssen sie fähig sein, sich unterschiedlichen Situationen flexibel anzupassen.

Telefonische Surveys

Mit zunehmender Telefondichte wurde die Möglichkeit, die Datenerhebung über das Telefon durchzuführen, immer häufiger genutzt. Telefoninterviews erfordern bei kleineren Stichproben keine großen infrastrukturellen Voraussetzungen und können relativ kostengünstig durchgeführt werden. Wenn die Telefondichte in einer Bevölkerung hoch genug ist, besteht darüber hinaus die Möglichkeit, große und/oder repräsentative Stichproben zu ziehen, sodass das Telefon sowohl für umfang-

reichere Surveys als auch für kleinere Erhebungen genutzt werden kann. Etwa seit der Jahrtausendwende sinkt die Zahl der Festnetzanschlüsse in den meisten westlichen Industrienationen jedoch zugunsten der Mobiltelefone kontinuierlich ab, sodass eine Repräsentativität von Telefonbefragungen über Festnetzanschlüsse schwieriger zu erzielen ist. In Deutschland waren 2014 zwar nur 10 % der Haushalte ausschließlich über das Mobiltelefon erreichbar, jedoch befinden sich darunter überproportional viele Jüngere (21 % in der Altersgruppe der 18- bis 29-Jährigen) und Arbeitslose. Darüber steigen die Verweigerungsraten bei Telefonbefragungen kontinuierlich an, da sich Telefonbesitzer durch die immer häufigeren Umfragen und Werbeanrufe belästigt fühlen. Nach einer Übersicht aus den USA sanken die Antwortraten bei Surveys mit kurzer Laufzeit (5 Tage) zwischen 1997 und 2003 von 36 % auf 25 %, bei aufwändigeren Studiendesigns mit mehrfacher Kontaktaufnahme von 61 % auf 50 %.

Aus Sicht der Untersucher sind telefonische Befragungen von Vorteil, da hier im Vergleich zur persönlichen Befragung die Wege- und Reisekosten wegfallen. Dadurch können in einer bestimmten Zeiteinheit wesentlich mehr Interviews durchgeführt werden. In größeren Studien kann der Einsatz von InterviewerInnen zentral über ein Surveylabor organisiert werden. Oftmals werden die Interviews dann als *Computer-Assisted Telephone Interviews* (CATI) durchgeführt. Hierdurch sind eine bessere Kontrolle der Studiendurchführung sowie eine bessere Supervision seitens der InterviewerInnen möglich. Ein weiterer Vorteil ist, dass Befragte bei Umzügen nicht mehr aus der Stichprobe ausscheiden, sofern sie ihre Telefonnummer beibehalten.

Andererseits muss das Studiendesign bei Telefonsurveys dem Medium angepasst werden. Zur Übermittlung von Informationen steht hier – zumindest bis heute – nur das gesprochene Wort zur Verfügung. Fragebögen, die für ein persönliches oder für ein schriftliches Interview konzipiert wurden, können daher für die telefonische Befragung untauglich sein. Sie sind möglicherweise zu komplex oder verwenden optische Präsentationen, wie z. B. Bilder oder Tabellen. In diesen Fällen muss eine Vereinfachung bzw. Adaptation vorgenommen werden. Wegen der begrenzten Gedächtnisspanne der telefonisch Befragten ist eine Präsentation von Antwortskalen oder längeren Listenfragen nicht möglich. Alternativ müssen Fragen zerlegt und die vorgegebenen Antworten in kategoriale bzw. in Ja/Nein-Formate transformiert werden. Da es bei reinen Telefoninterviews nicht möglich ist, zusätzliches Stimulusmaterial wie Fotos, Karten oder visuelle Hilfen zu verwenden, kann alternativ ein zweistufiges Verfahren gewählt werden. Hierbei wird zunächst der Kontakt zu den Befragten aufgebaut und erst nach der Zusendung dieses Materials dann das eigentliche Telefoninterview durchgeführt.

Bei Telefoninterviews ist die Latenzzeit zwischen Frage und Antwort kürzer als bei anderen Befragungsformen. Bei komplexeren Inhalten sowie bei Fragen, bei denen die Befragten auf ihre Erinnerungen zurückgreifen müssen, kann das zu einer vergleichsweise niedrigen Zuverlässigkeit (*Reliabilität*) führen. Auch ist die Art des Kontakts am Telefon anonymer als bei einer persönlichen Befragung. Andererseits

kann ein höherer Grad an Anonymität bei sensiblen Themen (wie z. B. beim Thema „häusliche Gewalt“) eine Befragung erst möglich machen.

Schriftliche Befragung

Bei schriftlichen Befragungen werden die Fragebögen per Post verschickt oder auf eine andere Art ausgeteilt. Dabei muss darauf geachtet werden, dass kein direkter Kontakt zwischen ForscherInnen und Befragten während des Ausfüllens besteht. Die Rücklaufquoten können stark zwischen 10 % und 90 % schwanken. Dies ist u. a. durch unterschiedliche Merkmale der Zielgruppen erklärbar. So sind Bevölkerungssurveys, die ohne ein offensichtliches Schwerpunktthema durchgeführt werden, anfälliger für eine geringe Rücklaufquote als thematisch enger definierte Befragungen. Auch bei bestimmten Bevölkerungsgruppen (u. a. bei Menschen mit hohem Zeitdruck, wie etwa Personen, die Beruf und Familie miteinander vereinbaren müssen) muss mit niedrigeren Beteiligungen gerechnet werden. Hohe Rücklaufquoten von über 70 % können v. a. dann erreicht werden, wenn bei den Befragten eine hohe persönliche Betroffenheit vorliegt (z. B. bei PatientInnen), wenn sie sich von der Teilnahme einen positiven Nutzen versprechen oder wenn ihnen die durchführende Institution bekannt ist.

Ein Programm zur Steigerung des Rücklaufs bei schriftlichen Befragungen („The Taylored Design Method [TDM]“ von Dillman et al.) beinhaltet die folgenden Maßnahmen:

- *Fragebogen*: Der Fragebogen sollte als gebundenes, ansprechend gestaltetes Heft konstruiert werden. Er sollte mit einem interessanten, aber neutral gestalteten Umschlag aus festerem Papier versehen sein. Die optimale Länge eines Fragebogens wird in der Literatur mit 12 bis 16 Seiten angegeben.
- *Anreize*: Die Befragten sollten eine kleine Anerkennung (keine Bezahlung!) für das Ausfüllen des Fragebogens erhalten. Verschiedene Studien konnten zeigen, dass dadurch auch die Bereitschaft zur Teilnahme an Wiederholungsbefragungen steigt.
- *Mehrfache Kontaktaufnahme*: Um die Rücklaufquoten zu erhöhen, sollten wenn nötig insgesamt vier Kontaktaufnahmen vorgesehen werden (erste Versendung und drei Erinnerungen).
- *Frankierter Rückumschlag*: Um die Bearbeitung des Fragebogens für die Befragten so einfach wie möglich zu machen, sollte jeweils ein frankierter Rückumschlag beigelegt werden.
- *Anerkannte Autorität*: Dem Fragebogen sollte neben einem personalisierten Anschreiben ein unterstützender Begleitbrief einer anerkannten Autorität beiliegen. Diese Persönlichkeit sollte im Hinblick auf ihre soziale Anerkennung in der Gruppe der Befragten sorgfältig ausgewählt werden und einen Bezug zur Thematik der Studie haben.

Internetsurveys

Aufgrund der zunehmenden Verbreitung des Internets wird dieses Medium immer häufiger auch für *Gesundheitssurveys* genutzt. Die elektronische Aufbereitung erlaubt es, z. B. Bilder, Filme und andere Medien flexibel einzubinden. Innerhalb des Surveys kann vielfältiges Material präsentiert werden, das den Befragten die Beurteilung eines Sachverhaltes erleichtert oder ihre Erinnerung unterstützt.

Die Verwendung von Onlinebefragungen hat in den letzten Jahren deutlich zugenommen. Parallel dazu wurden Studien durchgeführt, um die Möglichkeiten und Grenzen von Internetbefragungen auszuloten. Sie zeigen, dass sich InternetnutzerInnen und NichtnutzerInnen bisher hauptsächlich durch ihr Alter unterscheiden. Bei vorhandenem Internetzugang gibt es darüber hinaus Unterschiede in der Vertrautheit der NutzerInnen mit dem Medium. Dies wirkt sich auf die Erreichbarkeit von Zielgruppen aus. Das Hauptproblem von Internetsurveys ist jedoch die geringe Teilnahmebereitschaft. Erschwerend kommt hinzu, dass Internetsurveys während der Befragung sehr häufig abgebrochen werden. Auch wenn es Befragten möglich ist, zwischen verschiedenen Interview-Formen zu wählen, ist die Beteiligung in Internetsurveys durchgehend niedriger als in allen anderen Verfahren der Datensammlung.

Die zunehmende Verbreitung von Internetbefragungen wird daher die Teilnahmequoten, die in den letzten Jahrzehnten kontinuierlich zurückgingen, weiter sinken lassen. Zur Qualität der auf diese Weise gewonnenen Daten gibt es bisher nur wenige Studien. Die Aussagekraft der Daten muss jedoch durch eine niedrige *Responserate* (d. h. bei einer niedrigen Beteiligung) nicht notwendigerweise beeinträchtigt sein. Allerdings sind die bisher vorliegenden Untersuchungen zur Datenqualität von Internetbefragungen selektiv und decken nur einzelne Themen ab. Die Befunde sollten daher nicht verallgemeinert werden. Derzeit konzentriert man sich bei der Weiterentwicklung von Internetbefragungen auf das größte Problem, die Verbesserung der Antwortbereitschaft. In diesem Zusammenhang stellen sich auch neue Fragen im Hinblick auf vertrauensbildende Maßnahmen, insbesondere auch zu Transparenz und Datensicherheit. Im Vordergrund stehen dabei technische Fragen (z. B. nach neuen Verschlüsselungsverfahren) und Fragen zur Manipulierbarkeit der Daten.

Delphi-Befragungen

Die Delphi-Methode wurde entwickelt, um zu einem bestimmten Thema die Expertenmeinungen zu bündeln und im Verlauf von mehreren Befragungsrunden zu einer konsensbasierten Meinung zu diesem Themen zu kommen. Mit Hilfe von Delphi-Befragungen ist es darüber hinaus möglich, in einer unübersichtlichen Entscheidungssituation zu einheitlichen und begründeten Schlussfolgerungen zu kommen oder anstehende Entwicklungen vorherzusagen. Dies kann z. B. dadurch geschehen, dass Forschungsbedarf definiert und relevante Themen priorisiert werden. Im

Rahmen einer Delphi-Befragung kann es notwendig sein, quantifizierende mit qualitativen Forschungsmethoden (s. Kap. 2.4.5) zu kombinieren. Man spricht daher von einem *Mehrmethodenansatz*. So können z. B. standardisierte Interviews, aber auch Gruppendiskussionen oder andere qualitative Ansätze genutzt werden. Oftmals werden im Rahmen eines Delphi-Verfahrens mehrere Befragungsrunden mit anonymisierten Rückmeldungen hintereinander geschaltet, um die Konsensbildung zu fördern und den sozialen Druck durch Meinungsführer zu neutralisieren.

Delphi-Verfahren werden auch in der Public-Health-Forschung zunehmend häufiger genutzt. Wegen der Expertenzentrierung und der mangelnden Treffsicherheit von Vorhersagen stehen sie jedoch auch immer öfter in der Kritik. Es hängt daher in erster Linie von der Zielsetzung eines Delphi-Prozesses ab, ob es sich um einen sinnvollen und erfolgreichen Ansatz handelt.

2.4.5 Qualitative Datenerhebungsverfahren

Mit Hilfe der bisher beschriebenen standardisierten Verfahren werden *quantitative Daten* gewonnen. Es ist jedoch auch möglich, *qualitative Verfahren* zur Datengewinnung einzusetzen. In Public Health kommen diese Verfahren z. B. zur Anwendung, wenn es gilt, verständliche Fragen für Fragebögen zu entwickeln und diese dann mit Hilfe von *Fokusgruppen* (s. u.) zu testen. Qualitative Verfahren können auch dazu dienen, die in quantitativen Befragungen erzielten Erkenntnisse durch detailliertere Informationen zu ergänzen (z. B. mit Hilfe von *episodischen* oder *fokussierten* Interviews bzw. *Fallstudien*, s. u.). Wegen des erheblich größeren Zeit- und Personalaufwands bei der Erhebung und Auswertung dieser Daten können solche Methoden nur bei relativ kleinen Fallzahlen eingesetzt werden. Am häufigsten werden die folgenden qualitativen Verfahren angewandt:

Narrative Interviews

Die Befragten werden zuerst über die Modalitäten des Vorgehens informiert. Anschließend werden sie aufgefordert, über ein zuvor festgelegtes Thema zu berichten. Ein solches Thema kann ein vergangenes Erlebnis sein, wie z. B. die eigene Krankengeschichte, die dann sowohl beschrieben als auch bewertet werden soll. Die Länge der Erzählphase wird durch die Befragten selbst bestimmt. Kommentierungen oder Nachfragen von Seiten der Interviewer sollen weitgehend unterbleiben.

Episodische Interviews

In episodischen Interviews werden die Befragten aufgefordert, über spezifische Situationen (z. B.: „In welcher Situation sind Sie sich zum ersten Mal ihrer ‚Gesundheit' bewusst geworden?") oder über Kategorien von Situationen (z. B.: „Wann ist für Sie ‚Gesundheit' wichtig?") zu berichten. Mit Hilfe episodischer Interviews wurden bei-

spielsweise Studien durchgeführt, die Zusammenhänge von kritischen Lebensereignissen und dem Auftreten spezifischer Erkrankungen aufdecken konnten. Grundlage für diese Art der Befragung ist ein Leitfaden, der die anzusprechenden Themen enthält. Sowohl die Auswahl als auch die Gewichtung der Themen wird jedoch den Befragten überlassen. Das Interview wird entweder aufgezeichnet oder in anderer Form protokolliert.

Fokussierte Interviews

Fokussierte Interviews haben vor allem das Ziel, Hypothesen zu testen. Grundlage ist wiederum ein Leitfaden. Er dient dazu, die für die Befragten bedeutsamen Aspekte eines Themas zu erfassen und ihre Reaktionen festzuhalten. Bei Fragen nach dem Gesundheitssystem kann dies z. B. bedeuten, dass Befragte erklären, was für sie zum Gesundheitssystem gehört. Aus den gewonnenen Informationen wird schließlich eine inhaltliche Synthese gebildet, die dann als Grundlage für die weitere Hypothesenbildung dient. Bei fokussierten Interviews gehört es zu den Aufgaben der InterviewerInnen, Fragestimuli zu setzen und die Befragten ggf. aufzufordern, ihre Aussagen zu präzisieren.

Fokusgruppeninterviews

In Fokusgruppeninterviews werden Gruppen von Personen zu einer vorher festgelegten Thematik befragt. Die Interviewer geben dabei das Thema vor und strukturieren die daraus entstehende Diskussion. Im Idealfall sollte die Gruppengröße bei acht bis 10 TeilnehmerInnen liegen. Solche Fokusgruppeninterviews können z. B. dazu dienen, die Bewohner eines Quartiers zu den gesundheitlichen Risiken und Ressourcen in ihrem Wohnumfeld zu befragen. Um einen geeigneten Kreis von Interview-TeilnehmerInnen zu gewinnen, kann es erforderlich sein, mit potentiellen TeilnehmerInnen Vorinterviews zu führen. Die Gruppeninterviews sollten auf Band aufgenommen und später in zwei Stufen aufbereitet werden. In einem ersten Schritt werden die Gespräche und Diskussionen mit Hilfe inhaltsanalytischer Techniken ausgewertet. Hierbei werden die Informationen unter Verwendung von zuvor festgelegten Interpretationsregeln und anhand von eigens erstellten Kategoriensystemen klassifiziert. In einem zweiten Schritt werden Diskussionsmuster herausgearbeitet, um z. B. Verzerrungseffekte durch Meinungsführer zu erkennen.

Bei der Interpretation der Daten muss immer berücksichtigt werden, dass es sich bei einem Fokusgruppeninterview um eine künstliche Situation handelt. Die TeilnehmerInnen wurden durch das Forschungsteam ausgewählt. Die beobachteten und registrierten Interaktionen müssen daher nicht den verbalen Reaktionen entsprechen, die in einem natürlichen Rahmen auftreten würden, sodass Übertragungen auf andere Umgebungsbedingungen mit Vorsicht durchgeführt werden müssen. Fokusgruppeninterviews können nicht nur als eigenständige Methode, sondern z. B. auch als Ergänzung zu Fragebogeninterviews eingesetzt werden. Sie können hier u. a. zu

einem detaillierteren Verständnis der mit Hilfe von geschlossenen Antwortvorgaben erhobenen Informationen führen.

Einzelfallstudien

Einzelfallstudien betrachten einen spezifischen Fall im Quer- oder im Längsschnitt. Voraussetzung hierfür ist die Annahme, dass der für die Untersuchung gewählte Fall in seinen relevanten Merkmalen typisch und damit auf andere Fälle übertragbar ist. Als Untersuchungsobjekte kommen neben Personen auch Gruppen, Institutionen oder Organisationsstrukturen in Frage. Beispiele hierfür wären etwa Schulen, Gewerkschaften oder GesundheitspolitikerInnen. Im Rahmen einer Einzelfallstudie werden dann unterschiedliche Arten von Daten mit dem Ziel gesammelt, ein möglichst vollständiges Bild im Hinblick auf die Fragestellung zu erhalten. So können etwa die Auswirkungen der Einführung von Fallpauschalen auf die alltäglichen Abläufe in einem Krankenhaus anhand einer kleinen Zahl solcher Einrichtungen untersucht werden. Dazu werden z. B. Daten aus den Patientenakten, die Verweildauern und andere routinemäßig erstellte Dokumente herangezogen. Betrachtet werden aber auch typische Interaktionsmuster innerhalb des Krankenhauses sowie Veränderungen bei den alltäglichen Handlungsabfolgen.

Fallstudien können dazu dienen, die Ergebnisse quantitativer Studien zu ergänzen, Hypothesen zu formulieren oder relevante Aspekte einer gegebenen Fragestellung möglichst vollständig auszuleuchten. Die Erkenntnisse aus einer oder wenigen Fallstudien lassen sich jedoch nicht generalisieren, sie können lediglich Unterschiede zwischen den gewählten Untersuchungseinheiten (z. B. Krankenhäusern, Gesundheitssystemen oder einmaligen Ereignissen) aufzeigen.

2.5 Gesundheitsökonomie

David Schwappach

Nicht alle gesundheitlichen Ziele und Maßnahmen, die grundsätzlich wünschenswert wären, sind auch finanzierbar. Die Frage, wie begrenzte Ressourcen eingesetzt werden sollen, ist eine zentrale Herausforderung für die Gesundheitssysteme weltweit. Um hier Antworten zu finden, wendet die *Gesundheitsökonomie* wirtschaftswissenschaftliche Theorien und Methoden auf das Gesundheitssystem an. Zu den wichtigsten Aufgaben gehört dabei die *Kosten-Nutzen-Bewertung* gesundheitsbezogener Leistungen.

Entscheidungen über die Verteilung von *Ressourcen* müssen auf allen Ebenen eines Gesundheitssystems getroffen werden. Wird beispielsweise ein neues Arzneimittel zur Behandlung des Diabetes mellitus entwickelt, muss entschieden werden, ob das neue Medikament in den Leistungskatalog der Krankenversicherung aufgenommen und bei einer Verordnung bezahlt wird. Auch zwischen den verschiedenen Bereichen des Gesundheitssystems müssen die Mittel verteilt werden, z. B. zwischen

den Gebieten der präventiven, kurativen und palliativen Medizin. Das Gesundheitssystem konkurriert darüber hinaus mit anderen gesellschaftlichen Sektoren um Ressourcen, wie etwa dem Bildungswesen. Solche Überlegungen können durchaus sinnvoll sein, wenn zum Beispiel mehr Gesundheit „produziert" werden könnte, wenn verstärkter in die Bildung der Bevölkerung als in die Therapie bestehender Erkrankungen investiert würde.

Neben anderen Kriterien, wie zum Beispiel ethischen Überlegungen (s. Kap. 1.6), kann das **Kosten-Nutzen-Verhältnis** gesundheitsbezogener Maßnahmen eine wichtige Information für Entscheidungsträger sein. Im Bereich der Gesundheitsökonomie werden daher mit Hilfe spezifischer Evaluationsstudien Kosten/Nutzen-Daten erstellt. Als *gesundheitsökonomische Evaluation* bezeichnet man die vergleichende Analyse verschiedener Handlungsmöglichkeiten anhand ihrer jeweiligen Kosten und Nutzen. Für die zu vergleichenden Alternativen (z. B. zwei verschiedene Medikamente zur Senkung des Bluthochdrucks) werden die gleichen Kosteneinheiten (z. B. „Euro") und die gleichen Nutzeneinheiten (z. B. „Blutdrucksenkung in mm Hg") verwendet. Auf diese Weise lassen sich verschiedene Alternativen anhand expliziter Kriterien miteinander vergleichen. Favorisiert wird dann jene Alternative, bei der für die Erzielung einer Nutzeneinheit der geringere Mitteleinsatz erforderlich ist.

2.5.1 Gesundheitsökonomische Studientypen

Im Bereich der Gesundheitsökonomie lassen sich vier verschiedene Studientypen unterscheiden:

- Kosten-Minimierungs-Analyse
- Kosten-Effektivitäts-Analyse
- Kosten-Nutzwert-Analyse
- Kosten-Nutzen-Analyse

Sie unterscheiden sich darin, wie der Nutzen gesundheitlicher Maßnahmen ausgedrückt wird. In der Darstellung der Kosten gibt es keinen grundsätzlichen Unterschied (Tab. 2.11).

Die **Kosten-Minimierungs-Analyse** (Cost-Minimization-Analysis, CMA) ist eine so genannte reduzierte gesundheitsökonomische Studie. Ihr liegt die wesentliche Annahme zugrunde, dass sich der Nutzen der bewerteten Alternativen nicht unterscheidet. Damit reduziert sich die Analyse auf einen reinen Kostenvergleich. Die Annahme eines identischen Nutzens ist jedoch nur in sehr wenigen Fällen wirklich erfüllt. Zum Beispiel haben viele Maßnahmen zwar ähnliche erwünschte Wirkungen, aber ein unterschiedliches Nebenwirkungsprofil. Ergebnisse einer CMA sind daher immer kritisch zu prüfen.

Tab. 2.11: Merkmale der verschiedenen gesundheitsökonomischen Studien.

	CMA	CEA	CUA	CBA
	Cost-Minimization-Analysis	Cost-Effectiveness- Analysis	Cost-Utility- Analysis	Cost-Benefit-Analysis
	Kosten-Minimierungs-Analyse	Kosten-Effektivitäts-Analyse	Kosten-Nutzwert-Analyse	Kosten-Nutzen-Analyse
Kosten	Monetäre Einheiten (z. B. €)	Monetäre Einheiten (z. B. €)	Monetäre Einheiten (z. B. €)	Monetäre Einheiten (z. B. €)
Nutzen	Wird als identisch angenommen und nicht berücksichtigt	Natürliche Einheiten (z. B. Senkung des Blutdrucks in mm Hg; beschwerdefreie Tage)	Qualitätsadjustierte Lebensjahre (QALYs)	Monetäre Einheiten (z. B. €)

Die **Kosten-Effektivitäts-Analyse** (Cost-Effectiveness-Analysis, CEA) ist die am häufigsten eingesetzte Analyseform. Dabei wird der Nutzen in „natürlichen" Einheiten ausgedrückt, also solchen Parametern, die beobachtbar oder messbar sind. Beispiele für natürliche Nutzeneinheiten sind die Blutdrucksenkung in mm Hg, beschwerdefreie Tage oder vermiedene Frakturen bei Osteoporose.

Die Daten zum Nutzen der zu vergleichenden Maßnahmen werden entweder im Rahmen einer Studie neu erhoben oder aus der Literatur übernommen und anschließend ins Verhältnis zu den jeweils entstehenden Kosten gesetzt (s. eine Box in Kap. 2.5 auf unserer Lehrbuch-Homepage). So lässt sich für beide Maßnahmen ein **Kosten/Nutzen-Quotient** (auch *Kosten-Effektivitäts-Rate*, CER genannt) errechnen. Die CEA kann dann angewandt werden, wenn sich der Nutzen der zu vergleichenden Maßnahmen berechtigterweise im gleichen Nutzenmaß ausdrücken lässt. Die Reduktion auf nur einen spezifischen Nutzenaspekt wird den Konsequenzen vieler Erkrankungen aber nicht gerecht. Häufig haben sowohl Erkrankung als auch deren Behandlung Effekte auf die Lebensqualität und die Lebenserwartung der Patienten.

Die **Kosten-Nutzwert-Analyse** (Cost-Utility-Analysis, CUA) überwindet diesen Nachteil der CEA. Bei der CUA werden Wirkungen von Krankheit und gesundheitsbezogenen Maßnahmen auf Lebensqualität und Lebenslänge in einem „virtuellen", d. h. nicht real existierenden Nutzenmaß zusammengefasst, dem **qualitäts-adjustierten Lebensjahr** (Quality-Adjusted Life Year, QALY; s. Kap. 10.1.2). Als QALY bezeichnet man die mit einem Qualitätsfaktor gewichtete Dauer eines Gesundheitszustandes.

1 QALY ≙ 1 Lebensjahr in vollständiger Gesundheit

Es spiegelt die Annahme wider, dass ein Lebensjahr in guter Gesundheit für die Betroffenen einen höheren Wert hat als ein Lebensjahr in schlechter Gesundheit. Würde man also nur den Effekt von Maßnahmen auf die Lebenserwartung vergleichen, so blieben möglicherweise erhebliche Unterschiede in der Lebensqualität unbeachtet. Aus diesem Grund werden bei den QALYs die Lebensjahre durch eine Gewichtung „qualitätskorrigiert". Der Gewichtungsfaktor gibt den subjektiven Wert eines Gesundheitszustandes an und wird als **Nutzwert** (Utility) bezeichnet. Nutzwerte können einen Wert zwischen 0 und 1 annehmen. Der Wert 0 entspricht dabei dem Tod, der Wert 1 einem Zustand in vollständiger Gesundheit (Box 2.5.1, Abb. 2.21).

Box 2.5.1: Wie werden Nutzwerte ermittelt?

Nutzwerte können durch verschiedene Verfahren erhoben werden. Mit dem EQ-5D liegt ein standardisierter Fragebogen zur Erfassung von Nutzwerten in vielen Sprachen vor, bei dem sechs Dimensionen der Lebensqualität berücksichtigt und zu einem Wert zusammengefasst werden. Andere häufig eingesetzte Verfahren sind die *Time Trade-Off Methode* (TTO, Zeitausgleichsverfahren) und das *Standard Gamble* (SG, Standard-Lotterie-Verfahren). Für viele Erkrankungen und Gesundheitszustände existieren inzwischen publizierte Nutzwerte, die z. B. über die *Cost Effectiveness Registry* recherchiert werden können (Näheres zu den Verfahren s. Internet-Ressourcen).

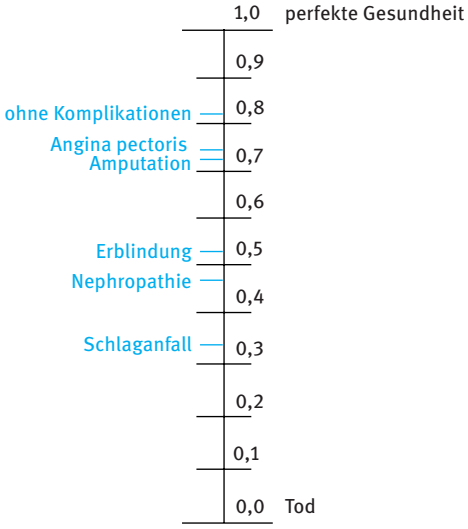

Abb. 2.21: Exemplarische Nutzwerte für verschiedene Komplikationen des Diabetes mellitus Typ 1; Abbildung basiert auf: Lee JM, Rhee K, O'grady MJ et al. Health Utilities for Children and Adults with Type 1 Diabetes. Medical Care 2011; 49 (10): 924–931.

QALYs berechnet man, indem man die Dauer eines Gesundheitszustandes (in Jahren) mit dem jeweiligen Nutzwert *multipliziert*. Verbringt also ein Patient fünf Jahre in einer Lebensqualität, die einem Nutzwert von 0,8 entspricht, so werden in dieser

Zeit vier QALYs angehäuft. Für eine CUA werden nun jeweils die QALYs der zu vergleichenden Maßnahmen berechnet. Diese werden dann ins Verhältnis zu den dafür aufzuwendenden Kosten gesetzt, sodass daraus ein *Kosten/QALY-Quotient* resultiert (Abb. 2.22).

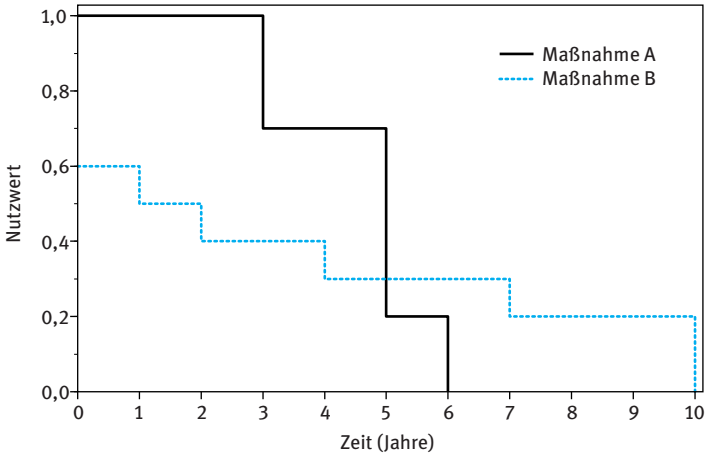

Abb. 2.22: Vergleich des Nutzens zweier hypothetischer Maßnahmen durch Berechnung von QALYs im Rahmen einer CUA. Bei Maßnahme A werden 4,6 QALYs angehäuft, bei Maßnahme B sind es 3,4 QALYs. Die Maßnahmen A und B verursachen Kosten in Höhe von 68.000 € bzw. 55.000 €. Damit liegen die Kosten/QALY-Quotienten bei 14.783 €/QALY für Maßnahme A und 16.176 €/QALY für Maßnahme B. Obwohl also die Patienten bei Maßnahme A kürzer leben und die Behandlung teurer ist, häufen sie mehr QALYs an und das Kosten/QALY-Verhältnis ist günstiger.

Der große Vorteil der CUA liegt darin, dass hier theoretisch alle Auswirkungen einer Maßnahme auf die Lebensqualität (z. B. physische, psychische und soziale Aspekte) und die Lebenslänge abgebildet werden. Damit wird das Kosten-Nutzen-Verhältnis gesundheitsbezogener Maßnahmen unabhängig von Indikationsgebieten, Versorgungssektoren und anderen Merkmalen miteinander vergleichbar. Hierin unterscheidet sich die CUA von der CEA, bei der dies durch das notwendige gemeinsame Nutzenmaß nicht möglich ist. Auf diese Weise könnten beispielsweise (1) ein neues Operationsverfahren mit höherer Überlebenswahrscheinlichkeit bei polytraumatisierten Patienten, (2) ein Screeningverfahren zur Früherkennung bösartiger Neubildungen bei Kindern und (3) eine medikamentöse Therapie bei chronisch Kranken anhand des Kosten/QALY-Quotienten miteinander verglichen werden.

Will eine Gesellschaft die Lebensqualität und Lebenslänge ihrer BürgerInnen maximieren, so sollte sie ihre Ressourcen auf die Maßnahmen konzentrieren, die das günstigste Kosten/QALY-Verhältnis aufweisen. Allerdings sollten die im Rahmen dieses Ansatzes der **„Nutzen-Maximierung“** auftretenden *ethischen Aspekte* kritisch reflektiert werden. Durch die Konstruktion des QALYs werden z. B. lebensrettende

Maßnahmen bei Kindern und Jugendlichen immer einen höheren Gewinn an QALYs aufweisen als lebensrettende Maßnahmen bei Erwachsenen, da durch die höhere Lebenserwartung jüngerer Menschen per Definition in der Zukunft mehr QALYs aufsummiert werden können (s. a. Kap. 1.4). Die CUA hat jedoch immer dann große Vorzüge gegenüber der CEA, wenn durch die betrachteten Erkrankungen oder Maßnahmen verschiedene Aspekte der Lebensqualität und/oder Lebenslänge beeinflusst werden.

Bei der **Kosten-Nutzen-Analyse** (Cost-Benefit-Analysis, CBA) werden nicht nur die Kosten, sondern auch die Nutzen in geldwerten Einheiten ausgedrückt. Dies bedeutet, dass alle gesundheitlichen Aspekte ebenso wie Todesfälle monetarisiert werden. Die CBA ist damit die einzige Analyseform, bei der sich ein „Netto-Nutzen" berechnen lässt (Nutzen [€] minus Kosten [€]). Damit können auch gesamtgesellschaftliche Mittelverwendungen verglichen werden (z. B. Vergleich zwischen Bildungs- und Gesundheitssystem).

Zur Bestimmung eines monetären Wertes von gesundheitlichen Effekten – wie z. B. dem Rückgang von Symptomen bei einer Erkrankung – werden häufig Verfahren zur Bestimmung der *Zahlungsbereitschaft* eingesetzt, die **Willingness-to-Pay (WTP)-** oder die Willingness-to-Accept (WTA)-Methode. Bei diesem Ansatz wird untersucht, wie viel Geld eine Person zu bezahlen bereit wäre, um einen gesundheitlichen Effekt zu erzielen bzw. wie viel Geld sie fordern würde, um auf eine Gesundheitsverbesserung zu verzichten. Für die Bestimmung des Geldwertes eines (vermiedenen) Todesfalles werden auch Ansätze aus der Versicherungs- und Verkehrsplanung sowie dem Arbeitsschutz verwendet. Die beschriebenen Verfahren zur Monetarisierung von Gesundheit und Lebenszeit werden kontrovers diskutiert. Zum einen wird hinterfragt, ob die geldwerte Bemessung eines menschlichen Lebens ethisch vertretbar ist, zum anderen können WTP- und WTA-Werte durch die Einkommensverhältnisse der befragten Personen beeinflusst werden. Die eigentliche CBA wird nur selten angewandt. Darüber hinaus wird der Begriff der Kosten-Nutzen-Analyse jedoch häufig auch als Oberbegriff für alle gesundheitsökonomischen Analysen verwendet.

2.5.2 Kostenarten

Bei gesundheitsökonomischen Evaluationen werden möglichst alle relevanten Ressourcenverbräuche berücksichtigt und anschließend mit den Kosten hinterlegt, die mit der zu untersuchenden Erkrankung oder der alternativen Maßnahme verbunden sind. Dafür ist es unerheblich, wer diese Kosten zu tragen hat (z. B. PatientIn, ArbeitgeberIn etc.). Eine Ausnahme bilden gesundheitsökonomische Studien, die explizit aus einer spezifischen **Perspektive** heraus durchgeführt werden. Oft ist dies die Perspektive des Krankenversicherungssystems. Bei solchen Studien werden dann nur jene Kosten berücksichtigt, die durch den Krankenversicherer getragen werden

müssen. Andere Kosten – z. B. solche, die die Patienten selber tragen – fallen heraus. Bei einer umfassenden, d. h. nicht aus einer bestimmten Perspektive unternommenen, gesundheitsökonomischen Studie werden folgende **Kostenarten** berücksichtigt:

Direkte Kosten

Medizinische und nicht-medizinische direkte Kosten sind alle Ressourcenverbräuche, die durch eine Krankheit und deren Behandlung entstehen. Zu den *medizinischen direkten Kosten* zählen beispielsweise die Kosten der Behandlung im Krankenhaus, von Medikamenten, diagnostischen Untersuchungen und Physiotherapiemaßnahmen. Auch die Behandlung von Nebenwirkungen ist hierbei zu berücksichtigen. *Nicht-medizinische direkte Kosten* sind z. B. die Fahrtkosten zum Krankenhaus, die Kosten von Umbauarbeiten im Wohnhaus eines Patienten aufgrund einer durch die Krankheit eingetretenen Behinderungen oder auch die Kosten, die durch die Inanspruchnahme von Hilfsleistungen im Haushalt von PatientInnen anfallen.

Indirekte Kosten

Viele Erkrankungen können dazu führen, dass Menschen in ihrer Produktivität eingeschränkt werden. Dies ist der Fall, wenn sie ihrer Erwerbsarbeit zeitweise nicht nachkommen können oder wenn sie dauerhaft arbeitsunfähig sind. Auch wenn sie frühzeitig versterben, entstehen *Produktivitätsverluste* („indirekte Kosten"). Produktivitätsverluste im Bereich der nicht-bezahlten Arbeit, z. B. der Familienarbeit, können ebenfalls relevant sein und müssen dann berücksichtigt werden. Gerade bei chronischen und psychischen Erkrankungen ist der Anteil der Kosten, der durch Produktivitätsverluste entsteht, oft hoch und kann damit einen erheblichen Effekt auf das Kosten-Nutzen-Verhältnis von Behandlungsmaßnahmen haben.

Intangible Kosten

Als *intangible Kosten* werden Kosten bezeichnet, die nicht oder nur schwer gemessen und dann in einem Geldwert ausgedrückt werden können. Beispiele hierfür sind Angst oder Stigmata, die mit einer Erkrankung verbunden sind. Intangible Kosten werden bei gesundheitsökonomischen Evaluationen nicht quantitativ berücksichtigt, zumal sie oft schon auf der Nutzen-Seite negativ in die Bewertung der Lebensqualität eingehen. So ist etwa davon auszugehen, dass PatientInnen, die unter einer besonders stigmatisierten Erkrankung leiden (z. B. psychische Erkrankungen, HIV/AIDS), deswegen auch in ihrer Lebensqualität eingeschränkt sind.

2.5.3 Die inkrementelle Betrachtungsweise bei gesundheitsökonomischen Studien

In den vorangegangenen Beispielen wurden die Ergebnisse gesundheitsökonomischer Studien als durchschnittliche Kosten/Nutzen-Quotienten für die zu vergleichenden Alternativen dargestellt. Unabhängig vom Studientyp ist die Verwendung von Durchschnitts-Quotienten jedoch oft irreführend und unrealistisch. Dies ist in der Medizin vor allem dann der Fall, wenn es um einen Vergleich mit neuen, oft teureren Alternativen zu bereits bestehenden Maßnahmen geht. In solchen Situationen ist vielmehr relevant, welche zusätzlichen Kosten aufgebracht werden müssen, um einen zusätzlichen Nutzen zu erzielen. Dieses Konzept wird als inkrementelle (schrittweise) oder marginale Betrachtungsweise bezeichnet (s.). Dabei wird nicht der durchschnittliche Kosten/Nutzen-Quotient einer Alternative berechnet, sondern der inkrementelle Kosten/Nutzen-Quotient (inkrementelle Kosten-Effektivitäts-Rate, ICER). Hierzu wird die Differenz der Kosten (Kosten der zu prüfenden Maßnahme abzüglich der Kosten der vorhandenen Alternative) zur Differenz der Nutzen beider Optionen ins Verhältnis gesetzt (Δ Kosten / Δ Nutzen s. Box 2.5.2).

Box 2.5.2: Fallbeispiel zur Errechnung des inkrementellen Kosten/Nutzen-Quotienten.

Die Patienten mit einer bestimmten psychischen Erkrankung erhalten üblicherweise über einen längeren Zeitraum eine effektive Verhaltenstherapie (Behandlung A). In einer Studie wurde nun ein neues Medikament untersucht, das zur Therapie dieser Erkrankung entwickelt wurde (Behandlung B). Folgende Berechnungen wurden vorgenommen:

Behandlung	Gesamtkosten	Gesamtnutzen	Kosten/Nutzen-Quotient	Inkrementeller Kosten/Nutzen-Quotient
A	425.000 €	18 QALYs	23.611 €/QALY	–
B	495.000 €	20 QALYs	24.750 €/QALY	35.000 €/QALY

Rechnung: Inkrementeller Kosten/Nutzen-Quotient = Gesamtkosten B – Gesamtkosten A / Gesamtnutzen B – Gesamtnutzen A

(495.000 € – 425.000 €) / (20 QALYs – 18 QALYs) = 35.000 €/QALY

Für jedes zusätzlich gewonnene QALY müssen bei der neuen Behandlung also 35.000 € mehr ausgegeben werden.

Häufig ist der inkrementelle Kosten/Nutzen-Quotient deutlich höher als der durchschnittliche Kosten/Nutzen-Quotient. Die inkrementelle Betrachtungsweise ist jedoch grundsätzlich vorzuziehen, da sie den Entscheidungsträgern realitätsnähere Informationen über den tatsächlichen Unterschied zwischen den betrachteten Alternativen aufzeigt.

2.5.4 Die Interpretation gesundheitsökonomischer Studienergebnisse

Die aus einer gesundheitsökonomischen Evaluation hervorgegangenen Ergebnisse lassen sich grafisch sehr gut mit Hilfe einer so genannten **Kosten-Effektivitäts-Fläche** darstellen. Prinzipiell sind dabei vier verschiedene Alternativen möglich (Abb. 2.23).

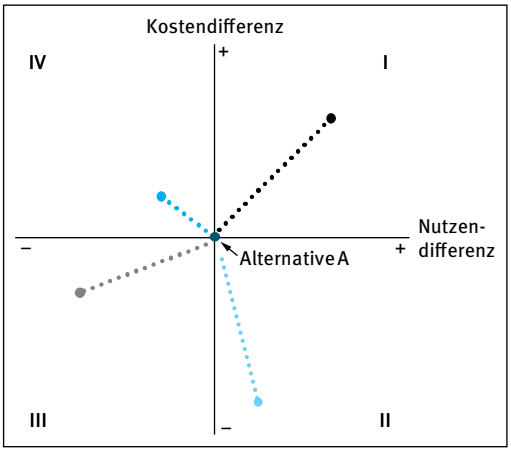

Abb. 2.23: Die Ergebnisse einer gesundheitsökonomischen Evaluationsstudie lassen sich auf einer Kosten-Effektivitäts-Fläche auftragen. Die untersuchten Maßnahmen werden im Vergleich zu einer Alternative A (im Zentrum; häufig der *Status Quo*) hinsichtlich der zusätzlich aufzuwendenden Kosten und des zu erwartenden zusätzlichen Nutzens eingetragen.

Im Vergleich zu einer Alternative A (im Zentrum) kann eine Maßnahme entweder teurer und effektiver (Quadrant I), günstiger und effektiver (Quadrant II), günstiger und weniger effektiv (Quadrant III) oder teurer und weniger effektiv sein (Quadrant IV). Da in den Quadranten II und IV jeweils eine der beiden Maßnahmen im Hinblick auf Kosten und Nutzen eindeutig über- oder unterlegen liegt, spricht man hier von einer **Dominanz**. Im I. Quadranten müssen für einen höheren Nutzen mehr Ressourcen eingesetzt werden, im III. Quadranten könnten dagegen bei Verzicht auf einen höheren Nutzen Ressourcen eingespart werden. In der Praxis tritt der positive inkrementelle Kosten/Nutzen-Quotient (Quadrant I) häufig auf. Er stellt Entscheidungsträger dann meist vor schwierige *Abwägungsentscheidung*.

Im Zentrum dieser Abwägungsentscheidung steht die grundlegende Frage der Gesundheitsökonomie: Wie viele Ressourcen ist eine Gesellschaft (z. B. ein solidarisch finanziertes Versicherungssystem) bereit, für eine zusätzliche Nutzeneinheit aufzuwenden? Wie viel darf beispielsweise ein zusätzlich gewonnenes QALY kosten? Bislang gibt es in Deutschland, Österreich und der Schweiz keinen eindeutigen, expli-

ziten Grenzwert zu dieser Frage. In Großbritannien existiert dagegen z. B. ein Schwellenwert von 30.000 £ (2016: ca. 33.000 €) bei der Zulassung neuer Medikamente. Er wird allerdings nie als alleiniges Kriterium angesehen. Stattdessen findet eine Abwägung verschiedenster, auch ethischer Kriterien, statt.

Ob der Einsatz von Ressourcen für eine bestimmte gesundheitsbezogene Intervention letztendlich „kosten-effektiv" ist, ist ein *bewertendes Urteil*. „Kosten-effektiv" ist dabei nicht gleichbedeutend mit „kosten-sparend". Da es sich immer um relative Berechnungen und Aussagen handelt, ist die Vergleichsgröße (z. B. die bisherige Behandlung) eindeutig anzugeben. Eine Intervention kann immer nur kosteneffektiv relativ zu einer anderen Alternative sein!

Internet-Ressourcen

Auf unserer Lehrbuch-Homepage **(www.public-health-kompakt.de)** finden Sie Literaturquellen, Hinweise auf weiterführende Literatur, zusätzliche Abbildungen und Tabellen sowie verschiedene Links zu interessanten Quellen.